Assessment, Equity and Language of Learning: Key Issues for Higher Education Selection in South Africa.

Nan Yeld

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Education, Faculty of Humanities

UNIVERSITY OF CAPE TOWN

August 2001

ACKNOWLEDGEMENTS

I would like to thank a number of people who have contributed to this study. I am especially grateful to my supervisors, Professors Joe Muller (Education) and Tim Dunne (Statistical Sciences), for their encouragement and careful consideration of the arguments and supporting evidence used in the study. I also wish to acknowledge the contribution of my colleagues in the Academic Development Programme and the Alternative Admissions Research Project. In this connection I am particularly mindful of my late colleague and friend, Jocelyn Hansen, whose support in the early years of the test development project was unstinting. Thanks are also due to Dr Martin Bygate of Leeds University, for his assistance in the early conceptualisation of the project. Finally, I wish to thank my partner, Robert Segall, and my children, for their patience and support.

ABSTRACT

The central problem investigated by this study arises from the fact that South African Senior Certificate results are not, for the majority of educationally disadvantaged candidates, reliable predictors of academic success in Higher Education. Despite this limitation, however, the Senior Certificate examination plays a vital role in the education system.

The aims of the study are thus to investigate procedures that could be used in addition to, rather than instead of, the Senior Certificate, and that would provide useful information about the future academic performance of educationally disadvantaged candidates. The purpose of these procedures is to widen effective access opportunities for such students.

It is clear that such procedures need to provide different information from that provided by the Senior Certificate which, like all achievement tests, aims to test learners' understandings in terms of the knowledge and skills covered in a preceding course of instruction. In contexts where great educational disparities exist, as is the case in the South African education system, it is neither fair nor defensible to base key gate-keeping events (such as entry to Higher Education) entirely on performance on such an examination.

Apart from issues of fairness, however, for students whose prior opportunities to learn have been grossly inadequate, achievement (curriculum-aligned) tests yield little useful information about candidates' underlying capacities and abilities. The study therefore investigates alternatives to achievement tests, and concludes that non curriculum-aligned testing of core skills and abilities could provide a workable alternative.

However, moving from curriculum-aligned to non curriculum-aligned tests can not in itself address the assessment challenge posed in identifying talented students in highly heterogeneous populations, in terms of educational preparation. In such contexts, educationally disadvantaged students will inevitably perform poorly in competition with their more advantaged peers, regardless of the basis of the tests. The study therefore reviews various approaches to what has become

known as dynamic assessment, and concludes that non curriculum-aligned, core skills tests developed as far as possible on dynamic lines may represent the most effective and fair approach to assessment in this context.

After reviewing major theories of knowing and learning, the roles of language in teaching and learning processes, and the history and possibilities of language testing, a set of specifications (a construct) is developed and proposed as the basis for an academic literacy test designed on dynamic lines.

The study then sets out to examine the Placement Tests in English for Educational Purposes (PTEEP), developed by the Alternative Admissions Research Project at the University of Cape Town. These tests aim to provide access opportunities for students whose Senior Certificate results do not necessarily reveal their potential to succeed at UCT. The investigation focuses on the extent to which the tests can be said to be (i) valid in terms of the construct established earlier, and (ii) useful in terms of providing useful, additional information about educationally disadvantaged candidates for selection purposes.

In other words, the first part of the study is devoted to developing, on the basis of an extensive literature review, a set of requirements for an academic literacy test for selection to Higher Education in South Africa. The second part of the study assesses the extent to which a series of tests developed by the author and currently being used for selection in this context, can be considered to be valid in terms of the construct established in part one.

Given the importance of English Second Language Higher Grade (ESL-HG) as the largest single subject registration in the Senior Certificate, and of English as language of learning, the study includes an investigation of the validity of the ESL-HG examinations, and of the usefulness of ESL-HG results for selection purposes.

The investigation employs both quantitative and qualitative research approaches. In summary, the analysis leads to the following major conclusions:

- overall, the PTEEP tests can be considered to be valid in terms of construct and content validity;
- the use of scaffolding within a test, for talented educationally disadvantaged candidates, can significantly enhance test performance;
- on the basis of survival analysis techniques (Polakow 1999), the PTEEP tests are effective in predicting academic success at UCT. That is, students who score in the top quintile of their candidate pool are significantly less likely to be excluded than are comparable students who are admitted on the basis of their Senior Certificate results alone. Students who score in the bottom quintile, however, have a very significantly higher risk of exclusion than their peers admitted on the basis of their Senior Certificate results alone;
- the PTEEP tests and the ESL-HG examinations exhibit divergent validity (that is, they are not positively associated, but reveal either random or inverse correlations); and
- ESL-HG and performance at UCT are not significantly associated.

On the basis of these conclusions, the study recommends that Higher Education institutions include, as part of their selection criteria and in addition to Senior Certificate results, a test that is non curriculum-aligned; based on the domain of academic literacy as defined in the study; and developed on the basis of dynamic principles. The study also recommends that the potential contribution of such a test to strengthen quality assurance at the school-leaving/Higher Education interface be investigated by the national Department of Education. Finally, it is recommended that, as a matter of urgency, the examining of ESL-HG be investigated, with particular reference to the extent to which the examination targets (and therefore contributes to promoting the development of) cognitive academic language proficiency.

LIST OF CONTENTS

Chapter Six:
Language Testing and the Assessment of Academic Literacy

Chapter Seven:
The PTEEP Test: Origins and History

Chapter Eight:
Construct Validity and the PTEEP Tests

LIST OF ACRONYMS

Advanced Progressive Matrices	APM
Alternative Admissions Research Project	AARP
American College Testing	ACT
Analysis of Variance	ANOVA
Basic Interpersonal Communicative Skills	BICS
Cognitive Academic Language Proficiency	CALP
Congress of South African Students	COSAS
Department of Education	DoE
Department of Education and Training	DET
Differential Item Functioning	DIF
English Language Proficiency Test	ELPT
English Second Language (Higher Grade)	ESL-HG
Further Education and Training Certificate	FETC
Historically Black Technikons	HBTs
Historically Black Universities	HBUs
Historically White Technikons	HWTs
Historically White Universities	HWUs
House of Assembly	HoA
Human Sciences Research Council	HSRC
International English Language Testing Service	IELTS
Joint Selection Project for Science and Applied Science	JSPSAS
Learning Potential Assessment Device	LPAD
Multiple Choice Question	MCQ
National Adult Literacy Survey	NALS
National Assessment of Educational Progress	NAEP
National Plan for Higher Education	NPHE
National Qualifications Framework	NQF
New Literacy Studies	NLS
Placement Test in English for Educational Purposes	PTEEP
Queensland Core Skills Test	QCS test
Raven Progressive Matrices	RPM
Senior Certificate	SC
South African Qualifications Authority	SAQA
Socio-Economic Status	SES
Technical and Further Education (Australia)	TAFE
Teach-Test-Teach	TTT
Tertiary Education Linkages Project	TELP
Test of English as a Foreign Language	TOEFL
University Foundation Year	UNIFY
University of Cape Town	UCT
University of the North	UNIN
University of the Western Cape	UWC
Vrije Universiteit Amsterdam	VUA
Zone of Actual Development	ZAD
Zone of Proximal Development	ZPD

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

AIMS, BACKGROUND, REASONS AND METHODOLOGY

1.1 Preamble

The central problem investigated by this study arises from the fact that South African Senior Certificate (SC) results are not, for the majority of educationally disadvantaged candidates, reliable predictors of academic success in Higher Education. At the same time, as is argued below, the crucial role played by the SC in the system means that it cannot simply be undermined or abandoned. It is clear that the SC must be improved and enhanced so that – amongst other important purposes - it yields reliable and useful information for Higher Education. In the meantime, however, it is necessary for the Higher Education system, or individual institutions, to consider the development of reliable and valid selection devices which can be used in addition to the SC, and which will provide the needed information.

The study investigates the possibility of constructing a selection device in light of the particular challenges in South Africa. It does so, as is explained in some detail below, by establishing theoretical and practical bases on which a selection device in such a context should be constructed, and then assessing the extent to which an existing, although hitherto not validated, testing initiative (for which the author is responsible) can be considered to be valid and reliable in those terms.

In Chapters Two to Six, relevant literature is reviewed with the aim of establishing the basis for an appropriate, effective and feasible selection device and procedure. The literature derives largely from the fields of Educational Assessment, Applied Language Studies, Language Testing, Educational and Cognitive Psychology, and Higher Education.

The second part of the study, Chapters Seven to Twelve, is empirical in nature, and takes the form of a validation study. That is, it focuses on the usefulness, for selection, of the Placement Tests in English for Educational Purposes (PTEEP) of the Alternative Admissions Research Project (AARP) at the University of Cape Town (UCT). These tests were developed by the author as leader of a development team, and are intended to provide an additional source of information in combination with information from the Senior Certificate examination. The tests are used by UCT and other institutions to provide access opportunities for students whose SC results would not necessarily reveal their ability to succeed in Higher Education.

In this part of the study, the standing of the tests in terms of various forms of validity, viz. construct, content, face, response, predictive, concurrent and consequential, is assessed. The data are derived from student performance at UCT over the years 1988 – 1998. The study also examines the English Second Language Higher Grade (ESL-HG) examination, the Senior Certificate subject with the highest registration. In this respect, a comparison is made of the effectiveness of the ESL-HG school-leaving examinations[1] and the PTEEP tests as predictors of future academic

[1] The study does not relate to other examination subjects in the Senior Certificate, but restricts itself to ESL-HG. Where data are available, however, the SC aggregate is included in the analysis.

performance in Higher Education. In addition to this comparison, the study investigates the possible reasons for - and implications of - any differences in such effectiveness.

In South Africa at present, the entire system of education, including Higher Education, is being radically overhauled in line with the transformation of the society. Key gate-keeping events, which have historically excluded black South Africans in particular, are under scrutiny, and the school-leaving examination, with its matriculation endorsement system (determining eligibility for university entrance), is one such event.

Recent and current developments in the school-leaving examination system, resulting from the elimination of racially based examination authorities, as well as from new Higher Education legislation, have proposed removing from the school-leaving examination the burden of gate-keeping for Higher Education, at the same time placing this function firmly at Higher Education's doorstep. These developments have created a climate in which the possibility of the establishment of a separate admissions testing system is being widely mooted. As yet, however, little that is concrete has been planned or achieved, and the need for a timely and comprehensively researched investigation of this area is clear.

Particular catalysts in the area include the following.

- The Higher Education Act (October 1997) proposes that eligibility for post-secondary educational opportunities be widened to include all who achieve the Further Education and Training Certificate (FETC), and that the matriculation endorsement mechanism[2], which effectively limited eligibility, fall away. Coupled with this widening is the requirement that Higher Education institutions develop systems of selection to manage the greatly increased pool of candidates eligible to apply for admission to Higher Education. One response to this requirement could be for the tertiary sector to

[2] Matriculation endorsement refers to the current minimum general admission requirements for Higher Education. The South African Certification Council (SAFCERT) issues a certificate attesting to (endorsing) a candidate's meeting a number of requirements relating to subject combinations, difficulty level and symbols attained (Zaaiman 1998:8).

establish an admissions testing system, independent of the schooling system, as has been done in other countries, most notably the United States.

If the FETC becomes the criterion for eligibility, the pool of potential Higher Education participants would be greatly increased. This increase would, in turn, force the issue of selection unavoidably onto the agenda even of Higher Education institutions, which have hitherto relied almost entirely on the matriculation endorsement mechanism as a selector. As access to Higher Education widens, moreover, so will the range – in terms of educational preparation - of students applying for admission. While it is relatively straightforward to select students at the top end of the school-leaving cohort, it is far more difficult to select amongst students whose school leaving examination performance is not as good. Methods of selection that can be used in addition to school leaving examination results are therefore needed. The study aims to identify such a method, or methods, and to assess its validity and usefulness.

Furthermore, the need to increase participation rates of black students, previously denied access to post-secondary education of high quality, means that ways of assessing academic potential need to be developed which go beyond simply confirming past experiences – what Miller (1990:52) calls "predicting the past". This consequence is particularly necessary now, as the previous divisions of education, while allowing and promoting discriminative practices, did, ironically, have the short-term advantage of ensuring that students were compared only with others with similar educational backgrounds. One of the consequences of the demise of the discriminatory but – in this sense - protective system has been the depression of the school-leaving examination scores of students at what were, and materially still are, Department of Education and Training (DET) schools[3]. This study aims to identify ways in which some of the principles involved in the testing of potential could be incorporated into national and/or institutional testing systems.

Current unease about the meaning of school-leaving examination results, in addition to recent legislative pressures, is in effect promoting the development of a national testing scheme which

[3] The DET system was designed to provide schooling for African students. It was severely under-resourced, in terms of funding as well as human capacity, with the majority of its teachers being under-qualified.

would operate outside of the school-leaving examination system. Such a scheme is already in place in a number of institutions, and unless strong links are maintained between it and the school-leaving examination, there is the danger that the school-leaving examination will lose much of its gate-keeping 'punch' (Yeld 1995, Yeld & van Bommel 1997). The study thus explores ways in which the contribution of a national testing scheme could be managed so that it does not disempower the school-leaving examination system, and aims to make recommendations in this regard.

Recent policy papers on Higher Education, such as the National Plan for Higher Education (Department of Education (DoE) 2001), the Green Paper (DoE 1996), and the Report of the National Commission on Higher Education (DoE 1997b), envisage growth in numbers[4] of students entering Higher Education of various kinds, and an increasing ease of articulation between different types of Higher Education provision. Both growth and increased mobility will highlight the necessity for institutions to develop flexible entry routes and curricula so that the needs and demands of 'non-traditional' students can be met effectively. The development of these flexible entry routes will, in turn, require reliable and accurate insights into the kinds of knowledge and skills brought by different groups of students, and required by different areas of study within various institutions. Such knowledge, essential for sound curriculum and course design, would be greatly facilitated through provision of information derived from well-designed admissions tests. Indeed, in this respect, the role of examinations in national education systems has increasingly come under the spotlight, for a variety of reasons. These reasons include a realisation on the part of many curriculum developers that innovations are unlikely to succeed if not reinforced in some way by the assessment system; that assessment is an essential tool in monitoring the effectiveness of such innovations; and that the power of progressive assessment practices to promote meaningful learning needs to be pursued.

The National Plan for Higher Education (NPHE), released in February 2001, reports that

> "[Current] poor graduation and retention rates and high drop-out rates are unacceptable and represent a huge waste of resources, both financial and human. For example, a

[4] At present the Higher Education system is experiencing a fall-off in demand, which, *inter alia*, mirrors the fall-off in the numbers of students obtaining Senior Certificate results which would make them eligible for consideration.

> student drop-out rate of 20% implies that about R1,3 billion in government subsidies is spent each year on students who do not complete their study programmes" (DoE 2001:22).

In order to address this problem, Higher Education institutions will be required to indicate, in the three year rolling plans on which their funding will be based, their "[S]election processes to determine the suitability of applicants who do not meet the minimum criteria for admission" (op cit:32). In addition, for both these students and for students who do meet the minimum criteria, institutions will have to indicate precisely how they intend to improve through-put rates: the main strategy proposed in this regard is the proper use of Academic Development Programmes.

In this scenario of heightened accountability and widened eligibility, the need is evident for a well-researched admissions procedure that will both widen access while ensuring a greater likelihood of success.

The issues sketched above are by no means new, but take on new urgency and complexity in times of great educational change such as that occurring globally in the late twentieth century, under the twin pressures of 'massification' of secondary and higher education, and the needs for increasingly sophisticated workforces, able to "... understand the technical systems they use and ... participate in dispersed management systems requiring judgement and decision making" (Resnick & Resnick 1992:38).

The seriousness of the challenge posed by these two pressures should not be underestimated. Education systems designed to cater for an elite are now required to 'deliver' to a vastly expanded clientele. Importantly, this expansion should be understood not only in terms of numbers, daunting though this is, but also in terms of an expansion and widening of needs in that many 'first-generation' learners in formal schooling bring with them widely differing literacy needs and backgrounds. In addition, however, education systems are now also required to provide opportunities to acquire more sophisticated skills, such as problem solving, understanding complex systems, and acquiring basic information technology skills.

Resnick and Resnick (1992), in a recent analysis of the United States schooling system, suggest that the 'routinised curriculum', which set out to teach basic skills (such as simple computation, reading of predictable texts) and was aimed at the huge majority not intending to proceed to Higher Education, is now not adequate for society's needs. They conclude that what is needed is a 'thinking curriculum' for all, not an elite few. The parallels with South Africa are striking: the system of Bantu Education (embodied in the DET), designed to ensure that African school-children would acquire only basic skills that would equip them for low level jobs has left a lingering legacy in the form of an under-skilled workforce and consequent low levels of production. Eradicating this legacy would be greatly facilitated through ensuring that higher-order cognitive skills are elicited at all levels of schooling.

Clearly, in light of these pressures, all aspects of the educational process need to be investigated to ensure that their potential to assist in the modernisation and effectiveness of the curriculum is maximised. As Kellaghan and Greaney (1992:65) point out: "... un-reformed examination systems are likely to cause damage to the quality of education offered in schools. It obviously makes sense to take steps to change such systems and to develop examinations which are more in keeping with principles of good assessment practice".

Assessment is thus a critically important factor in education. If poorly designed or inappropriately used, it can impede and distort the processes it was intended to serve; a point discussed in Chapter Three. If ignored, its potential to support and promote sound learning and teaching practices is lost.

As is evident from the aims of the study outlined above, the study targets the critical area of language in relation to learning. The focus on English is largely the result of the particular place of the language in South African society. Although less than 10% of the South African population speaks English as a first language, it is the medium of instruction in the great majority of the country's schools and Higher Education institutions (all 21 universities, for example, use English as a medium of instruction, and eight offer instruction also in Afrikaans). This depiction of the status quo should not, however, be taken to imply uncritical support for the hegemony of English in formal

education in South Africa. As will become clear in the course of the study, the selection device investigated is not necessarily specific to any language. For example, the PTEEP is currently used in translation, by a client institution whose language of learning is Afrikaans. The translated test is called the '*Plasingstoets in Afrikaans vir Opvoedkundige Doeleindes*'. Should other languages be used in the future as medium of instruction (for example, should the University of Zululand elect to teach and examine in isiZulu), there is no in principle reason why the PTEEP tests should not be translated into that language.

English Second Language occupies a unique place amongst languages in education in South Africa. It is a second (or additional) language for the great majority of the country's learners, and it is also, from a relatively early stage, the medium of instruction, or language of learning. In the time allotted for the teaching of language, however, it is treated like any other second language (e.g. Afrikaans Second Language, or isiZulu Second Language in KwaZulu Natal). So while an English first-language learner spends approximately six hours per week learning English, and taking her/his subjects through that medium, English Second Language learners spend approximately four hours learning English (Kapp 2000b) but still are required to use it as medium of instruction. The result is that learners who most need instruction in the language which is the medium of instruction, as it is not their first language, actually have less teaching and learning time set aside for this purpose than learners whose first language it is.

The PTEEP was developed by the AARP Project at UCT in recognition of the crucial role of the language of learning, English, in the academic progress of its students. The tests are based on the notion of what Cummins (e.g. 2000, 1984, 1980) has called 'cognitive academic language proficiency' (CALP), and aim to identify students who will succeed in their studies, while recognising the effects of educational disadvantage on test performance. The main focus of the second part of the study is directed at investigating the validity of the PTEEP tests.

The reasons for the analysis of the PTEEP tests include the following. First, the genesis of the AARP Project as a politico-educational strategy to increase the recruitment of black students meant that it was not originally guided by strong and professional assessment procedures and

principles. The need for professionalisation of the Project is now acute. This study sets out to assess the validity of the tests in the light of the data that are available, and to make recommendations on how the project might strengthen its test development procedures as well as its research capacity. Second, while recent statistical studies into the predictive validity of the tests have yielded promising indications about the value of the tests in identifying talented yet educationally disadvantaged students, a full and developmental validation study has not been conducted. The study sets out to rectify this by undertaking such a validation exercise. Finally, there is a real lack of properly researched local test development initiatives, and the comprehensive validation procedure undertaken in this study will, it is argued, make a valuable contribution in this regard.

In addition, the existence of the most widely used English language examination at this level, the English Second Language - Higher Grade (ESL-HG) examination, cannot be ignored. It is brought into the study mainly through the investigation of the concurrent validity of the PTEEP tests (that is, in Chapter Eleven). In addition, it features in the analysis of the predictive validity of the PTEEP (Chapter Ten).

ESL-HG is taken for the Senior Certificate examination by approximately 90% of South African Grade 12 learners. There are many reasons for investigating the way in which the subject is assessed. The most salient reason in the context of this study is the use of the ESL-HG examination for selection to Higher Education. The widespread belief that performance in the medium of instruction is an important indicator of academic performance at tertiary level persists in the face of much evidence that casts doubt on the strength of this relationship. For example, predictive studies of such internationally used tests of academic language proficiency as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) have shown that results on these tests explain only about 10% of the variance in subsequent academic performance of students (McNamara 1996)[5]. Reasons for this low predictive power are discussed in Chapter Ten, in particular.

[5] These studies refer to the language proficiency test and academic performance of students for whom English is a Second Language, and who are studying in a tertiary context where English is the medium of instruction.

A second powerful reason for the significance of ESL-HG in the SC examination set is that it is the largest subject registration, and for this reason alone it deserves close scrutiny. Ensuring that it is examined fairly and rigorously, and that it reflects theoretically and practically sound notions of language and language use would have major benefits for the maximum number of Senior Certificate candidates. Related to this objective is the potential of the ESL-HG examination to play a powerful role in the establishment and maintenance of a comprehensive quality promotion system. As the single largest subject, and on the basis of its language-across-the-curriculum role as the logical underpinner of CALP-type skills and abilities, it is ideally placed to act as a cross curriculum skills 'canary' for the system at the Grade 12 (end of schooling) level.

Finally, the examination is crucial because, as the embodiment of the syllabus, it shapes how the subject is taught and learned in the years preceding the examination. This power to influence classroom practice is usually thought of as a negative one, but it depends largely on the way the examination is conducted and designed and so forth. As Kapp (2000b:4) suggests, "[T]he matriculation question paper ... becomes a literacy practice in itself" – and, if it models a progressive and emancipatory practice, could be seen as a powerful tool for positive curriculum change. As it is, however, the apparent mismatch between the aims of the curriculum and the ways in which it is assessed (Hansen 1997) needs to be addressed as a matter of urgency.

The significance of language proficiency, and the contested area of how it should be defined in terms of its role in learning, is therefore of crucial importance to education in the country. Developing and implementing ways of ensuring that it is properly assessed is integral to making sure that it is taken seriously and assumes its rightful place in learning/teaching processes.

1.2. Aims and Research Questions

Several aims have been identified in the introductory remarks above. These are as follows:

- To investigate methods of selection that could be used in addition to SC results (i.e. where the objective is not to replicate information that can be gained from the SC, but to obtain different, complementary insights into the abilities of applicants);
- To identify ways in which access for educationally disadvantaged applicants could be widened, while at the same time minimising academic risk.
- To identify ways of testing potential, and of how these ways could be incorporated into national and/or institutional testing systems.
- To investigate the ways in which the ESL-HG is examined, and to make recommendations about how these can be strengthened and improved, so that this important area can more effectively promote the development of CALP-type skills.

The aims can be expressed in terms of two overarching research questions, as follows.

On what basis ought a selection test to Higher Education, in a context of widespread educational disadvantage, to be constructed? In other words, what would be an appropriate construct for such a test?

Chapters Two to Six of the study are devoted to investigating this question.

As was described above, however, the urgent need to increase the intake of black students at UCT led to the establishment of the AARP project in the late 1980s, and the consequent rapid development of selection tests. For reasons that will become clear in Chapter Seven in particular, these tests have not previously been comprehensively validated. Their continuing and increasing use by UCT and other South African Higher Education institutions, however, means that there is an urgent need for such a validation process in order to better understand the contributions of the tests to equitable and effective access, and to strengthen future development cycles.

The second major research question is addressed through a full and developmental validation study. The question is as follows:

To what extent are the PTEEP tests, developed to identify talented but educationally disadvantaged candidates whose SC results would not necessarily reveal their abilities, valid in terms of the construct and framework articulated above?

Chapters Seven to Twelve are devoted to a comprehensive validation study aimed at addressing this question.

1.3 Description of the Study

As stated above, the study is located in the field of assessment. Within this broad area, the focus is selection testing for Higher Education, and within that, the focus can be more narrowly defined to the role and contribution of language testing in selection for Higher Education.

The first part of the thesis, comprised of Chapters Two - Six, follows this progressive narrowing of focus. That is to say, Chapter Two focuses on selection issues and choices confronting Higher Education, and to particular concerns in terms of how student bodies are to be constituted. Chapter Three discusses, in broad detail, some of the main issues relating to assessment, and in particular to concerns about fairness and bias. Chapter Four focuses on the contribution of various assessment possibilities in selection to Higher Education. Chapters Five and Six examine in some detail the domain - academic literacy - that could form the basis for the development of appropriate tests for the purpose of providing qualitatively different but relevant information for use in selection to Higher Education.

Having established the major concerns and analytical frameworks for an analysis of language tests used in selection, the focus shifts, in Chapters Seven to Twelve, to a detailed analysis of the PTEEP tests, and a minor analysis of the ESL-HG examination, as these are two major language tests used in selection by the University of Cape Town.

A more comprehensive description of the study follows.

Chapter Two focuses on the area of selection to Higher Education. It lays out some of the main purposes and objectives of Higher Education, and the implications these purposes have for the

constitution of student bodies and the ways in which institutions can go about achieving their aims. In addition, the chapter contains a comprehensive discussion of general issues related to selection. Alternatives to selection are put forward and analysed. These alternatives include such options as raising tuition prices, random selection, 'fail-first' approaches, and so on. The discussion then moves to clarifying the factors that need to be taken into account in the development of selection criteria. These factors include such concerns as the desired composition of the student body, schooling exit levels, academic merit (and the need to predict academic performance), and institutional missions.

Chapter Three sets out to clarify and demystify several core issues related to assessment. The origins and characteristics of external, standardised testing are discussed, and related to the assessment environment in South Africa. In addition, the chapter addresses the issue of 'measurement-driven instruction', and the extent to which the positive potential of approaches within this paradigm can be harnessed for use in the South African context. The chapter also discusses various types of tests, such as achievement, proficiency, and aptitude tests. The chapter concludes with a discussion, in light of the analysis of selection-related issues, of the crucial importance of fairness in selection processes and procedures. The discussion covers such phenomena as adverse impact, discrimination and bias, and educational disadvantage, and analyses equal opportunity and affirmative action models within the constraints imposed by the need for effectiveness and efficiency in the Higher Education system.

Chapter Four moves from the general issues related to assessment covered in Chapter Three, to a discussion of the role/s assessment can play in selection for Higher Education. In other words, it focuses on a particular purpose for assessment, and on the options available in this regard. The pressures faced in developing countries in respect of assessment, where one examination commonly has to perform selection, monitoring and certification functions, are highlighted, and some of the ways in which these pressures can be managed are put forward. Moving to a review of selection studies in sub-Saharan Africa as well as elsewhere, some general problems, such as truncated samples, small sample sizes, and difficulties with defining the criterion, are outlined. In

addition, Chapter Four discusses two major approaches to selection practice and related research. These approaches can be characterised as static or dynamic methods of assessment. The merits and demerits of these methods are discussed in relation to the South African context, and their implications for future directions are spelled out.

Chapter Five, which takes as a starting point the need for selection procedures to go beyond sole reliance on achievement tests, such as the Senior Certificate, focuses on identification of an appropriate basis for such testing. That is to say, while Chapter Four argues the case for the development of tests which go beyond simply assessing what students have been taught, which, for many students, is an unfair, ineffective and inappropriate approach, Chapter Five investigates and establishes what it is that should be tested. In doing so, it explores the testing of competencies in core, non subject-specific skill areas as a basis for the development of tests that aim to go beyond achievement testing. Examples of relevant initiatives elsewhere are briefly examined, as are local initiatives such as the identification of critical cross-field outcomes by the South African Qualifications authority as essential to underpin all qualifications registered on the National Qualifications Framework. The core cognitive skills area identified and explored in Chapter Five as a basis for non-curriculum aligned assessment is that of academic literacy. More specifically, the discussion centres on literacy, and academic literacy, and the assumptions about knowing and learning which shape these. Behaviourist, differential, cognitive and situative perspectives on knowing and learning are reviewed in relation to literacy and academic literacy, and in each case the major implications for assessment are examined and their significance for use in selection procedures explored.

Chapter Six turns to an examination of the roles of language in knowing and learning, and on the implications of these roles for the assessment of academic literacy. As was highlighted in Section 1.2 above, this area is critically important in South Africa. In essence, Chapter Six contains a discussion of language and language testing in relation to developing and assessing academic literacy. That is, it includes a discussion of the history of language testing, which reflects the changing conceptions of what it is to know and use a language put forward in Chapter Five. It also

includes an analysis of challenges and possibilities for the assessment of academic literacy in a context of widespread educational disadvantage. The chapter concludes by articulating a construct for a text of academic literacy in such a context.

The second part of the study is comprised of Chapters Seven to Twelve. It focuses on the academic literacy tests developed by the Alternative Admissions Research Project - the PTEEP tests. In this process, an assessment is made of the extent to which claims can be made that the PTEEP tests are grounded in principles of good assessment practices, sound and appropriate theories of knowing and learning as applied to assessment possibilities, and are responsive to and appropriate in a context of widespread educational disadvantage. In addition, recommendations are made about the possible use of the tests in access and selection to Higher Education, in combination with the SC.

Chapter Seven lays the foundations for the second part of the study. It deals with the origins and history of the Alternative Admissions Research Project at UCT, with particular reference to the PTEEP tests. Test rationale issues such as test purpose, test users, test takers, and resources and constraints, and particular testing challenges faced by the Project, are discussed in this chapter. The main point of departure is the shaping influence these factors have had on the development of the test.

Chapter Eight focuses on construct validity: a central, overarching concern in test development. The chapter sets out to assess the extent to which the PTEEP tests can be said to be valid in terms of the general construct outlined at the end of Chapter Six. The analysis draws on various sources and kinds of evidence, both quantitative and qualitative, to test the claim of construct validity.

Chapter Nine focuses on content validity: that is, it assesses the extent to which what is in the tests (texts, tasks, items) can be said to represent the set of principles or requirements laid out in the construct. In addition to content validity, Chapter Nine focuses on face and response validity: that is, what a test looks like, what it appears to measure, and how it is experienced by the test taker.

Chapter Ten investigates predictive validity. For selection tests, the primary question is whether or not the tests succeed in providing a basis for successful selection decisions. The chapter discusses the predictive validity of the PTEEP tests in some detail, drawing on the findings of various statistical studies, and investigates the success of the PTEEP tests in terms of this purpose.

Chapter Eleven contains an analysis of concurrent and consequential validity. In investigating the concurrent validity of the PTEEP tests, the ESL-HG examinations and syllabuses are subjected to a similar, albeit abbreviated, analysis to that undertaken in the previous chapters. In relation to consequential validity, the positive and negative consequences of the establishment and introduction of the PTEEP tests are examined.

Chapter Twelve addresses the issue of reliability, and the discussion illustrates some of the difficulties caused by the testing approach in relation to the employment of classical test validation techniques.

Chapter Thirteen, the third part of the study, contains the conclusions and recommendations emanating from the study. These recommendations refer not only to the development of an appropriate selection device, but also to its relationship with the Senior Certificate examination and its likely successor, the FETC.

1.4 Scope and Limitations

Several limitations constrain the conclusions and generalisations that can be drawn. For example, the data relate to the performance of students from one type of educational background, at one Higher Education institution. In addition, the study draws on findings from a project set up initially as an admissions project rather than as an assessment project. At times, this focus on admissions led to somewhat opportunistic and unorthodox methods of test development and validation, and the difficulties these cause for analysis are demonstrated and discussed in the empirical part of the study. Nevertheless, preliminary indications are that the findings of the study will be sufficiently

robust to allow the making of several recommendations concerning the development of selection devices for Higher Education.

1.4.1 Student Performance Data

The student data on which the analyses in Chapters Eight to Twelve are based are derived from the performance of ex-DET students at UCT over the years 1988 – 1998. All of the students in the study had attended schools that were under the authority of the DET, or would have been so categorised prior to 1996 when the DET ceased to exist.

Various measures of performance for the years 1988, and 1995 –1999 were used in the study, as follows:

- Academic results at UCT. University performance was characterised in terms of three Indices:
  Index 1: the actual mark (expressed as a percentage) obtained by a student for each of her/his courses, weighted in terms of time and degree points;
  Index II: performance on courses in relation to the number of courses a candidate has taken; and
  Index III: the readmission status of a student in terms of UCT regulations.
- Performance on the PTEEP tests.
- Performance on the SC examination. SC performance was based on raw aggregate scores, scores derived from the UCT admissions points system, and ESL-HG results.

1.4.2 Documentary Data

The documents used in the analysis and discussion in Chapters Eight to Twelve are:

- PTEEP and English Language Proficiency Test (ELPT) documents: the tests themselves, test development documents (records of test development meetings, etc.), and reports written about the tests;

- AARP Project documents: Annual Reports, student records (biographical and academic performance), university documents (e.g. planning documents relating to the establishment of the Project), and school-leaving examination results;
- ESL-HG syllabuses from the years 1986 - 1995;
- A selection of DET ESL-HG examination papers from the years 1994 and 1995, and of the 1996 Western Province ESL-HG papers; and
- Reference works relating to language testing.

1.4.3 Assumptions of Equivalence Underlying the Data

The AARP tests under scrutiny in the study are, as described above, the ELPT and the four PTEEP tests developed and used for the 1988 - 1999 intake years. However, item-level data have only been recorded in the Project since 1998 (until this time only test totals were routinely entered), and thus such detailed data are only available for the 1998 and 1999 tests.

According to Nitko (2001), parallel forms of tests should be built on the basis of the same blueprint (the same specifications); have equal observed score means and standard deviations; have equal standard errors of measurement; and should correlate equally with other measurements. Equivalent tests, on the other hand, need to fulfil less stringent requirements. Although they must be "... based on the same specifications, they may vary as to the number of items, the response types and the content.... What is important ... is that they each measure the same language skills, and that they correlate highly with each other" (Alderson, Clapham & Wall 1995:288).

The PTEEP tests, as is demonstrated below, meet many of these criteria. Each test has not, however, been correlated with its successor, as it has not been possible to require candidates to write two tests. In seeking to establish equivalence, the following sources of evidence are advanced (1.4.3.1 - 1.4.3.3 below).

1.4.3.1 Test Specifications

As is discussed particularly in Chapter Nine, which addresses the content validity of the PTEEP tests, all of the tests are based on the same set of specifications, and care is taken to ensure even

coverage for each test, as is illustrated in Table 9.2 in that chapter. Explicit test development records document this subjective claim, to render it accountable if not yet testable.

1.4.3.2 Candidate Pool

It is argued that numbers of candidates are large enough to allow the assumption that the candidates are equivalent, in that over 1,500 candidates with similar educational backgrounds (ex-DET) wrote the tests each year, at the same time of year. The assumption of equivalence in this respect effectively infers that year-to-year cohort differences have not yet emerged in consequential ways to impact upon PTEEP testing and scores.

1.4.3.3 Use of Scores

It has been the practice since the early days of the Project to recommend the top 20% of each category of candidates (the categories were derived on the basis of educational background) for admission to the institution. There has thus been no development of a cut-off score. The absence of a defensible cut-off score means, of course, that a candidate's standing, or ranking, has been dependent on the ability level of the candidate pool. Several difficulties can arise in this regard, the main one being that applicants over the years might differ in terms of ability or preparation. For this reason, no admission recommendations are made until after all scripts from the first testing session for a particular intake year have been marked. As the great majority of candidates (for any year) write at this first session, it is believed that this ensures a large enough sample on which to derive quintile boundaries. As Figure 1.1 below demonstrates, there has been considerable stability over the years in terms of the quintile boundaries, and thus it can be argued that, for the purposes of selection, the tests could be regarded as equivalent.

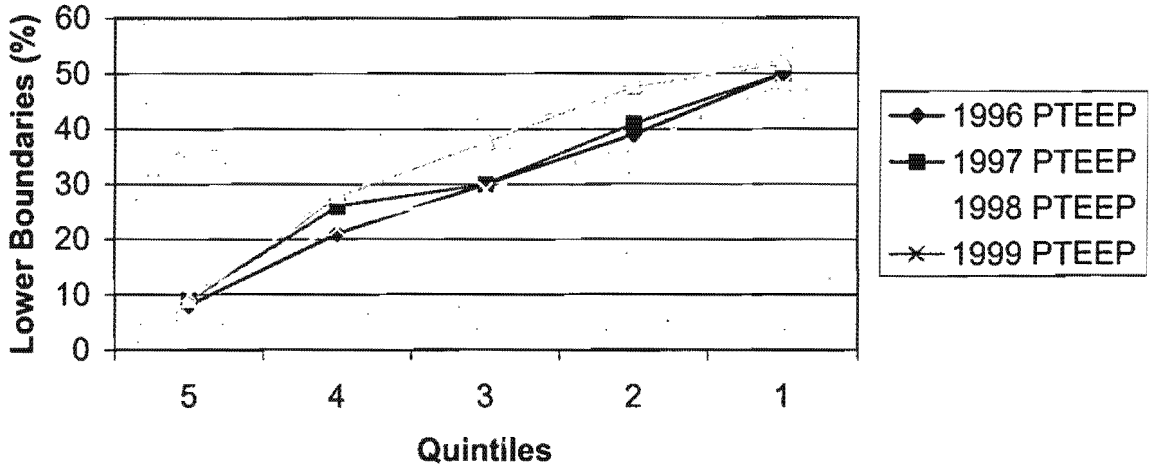

Figure 1.1: Quintile Boundaries

It can be seen from Figure 1.1 that the observed quintile boundaries are similar, particularly in connection with the lower boundary of the top quintile (Quintile 1). The 1998 intake test stands out somewhat as possibly being easier, except, again, at the lower boundary of the top quintile. This exception is crucial, as it is candidates in the top quintile who are recommended for admissions, and it is thus the university performance of the top quintile candidates that will determine the confidence with which admissions officers can use the Project recommendations. The table of percentage scores at the quintile boundaries from which this graph is derived is given below:

QUINTILES	1996 PTEEP	1997 PTEEP	1998 PTEEP	1999 PTEEP
1	50	50	53	48
2	39	41	48	37
3	30	30	38	29
4	21	26	28	22
5	8	9	9	9
No. of Candidates	1,153	883	915	959

Table 1.1: Lower Quintile Boundaries (%) of ex-DET candidates

Table 1.2 below contains descriptive statistics of the four PTEEP tests. The data used for the 1996 and 1997 intake years refer only to the performance of students who had written the PTEEP and had registered at UCT. Some students in this category would not have been recommended by the Project, but since the institution accepted students on the basis of their Senior Certificate results and/or their PTEEP results, they were admitted despite poor performance on the PTEEP. The 1996 and 1997 intakes are, however, a more select group than those whose performance is

reflected in the 1998 - 1999 intake tests, as the data from the latter years include all applicants (as opposed to registered students only). This contrast explains why the number of candidates is smaller in 1996 and 1997 than the other two years.

	Registered Students		Applicants	
	1996 PTEEP	1997 PTEEP	1998 PTEEP	1999 PTEEP
Mean	50.62	50.23	41.96	33.55
Std error	0.61	0.75	0.50	0.53
Std dev	10.64	10.37	15.27	16.42
No. of Candidates	302	193	915	959

Table 1.2: Descriptive Statistics, PTEEP Tests 1996 – 1999

Table 1.2 shows some similarities between the means, standard errors, and standard deviations of the two sets of tests: that is, the 1996 and 1997 tests, and the 1998 and 1999 tests. The 1998 and 1999 tests, however, differed considerably in their means, with the 1999 test being more difficult.

On the basis of these statistics, it can be argued that the tests appear sufficiently equivalent to perform similarly in respect of predictive, concurrent and consequential (i.e. external) validity. In terms of internal aspects of validity (construct, content, face, response), claims to equivalence rest on more qualitative sources of evidence, and these are discussed in the relevant chapters below.

1.5 Methodology

The study employs a variety of methodological approaches, drawing on both quantitative and qualitative paradigms. This eclectic approach is made necessary by the wide range of sources on which the study draws, as can be seen in 1.4.1 and 1.4.2 above.

The methods used to interrogate the documentary sources are discussed in more detail in Chapters Eight to Twelve, in the context of the specific analysis undertaken in each chapter. For example, the analysis of content validity and the analysis of concurrent validity require very different analytical approaches, and these are best described and motivated in the process of these validation undertakings. A somewhat more extended account (1.5.2 below) is provided of the statistical techniques employed, however, as they are used throughout Chapters Eight and

Twelve, and the conditions which govern their appropriate use can be more easily abstracted from a particular context.

1.5.1 Qualitative Approaches

The documentary data illustrated in 1.4.2 above are of several different types, and require different analytical approaches. One type is comprised of documents about the Project – its history and origins, and progress over the years. The methods used to analyse these are similar to those used in historical research, which essentially aims to "... use the past to understand and explain the present ..." (Cohen & Manion 1994:49). As was stated above, the origins of the AARP project have had a significant impact on the development of the tests, and a major aim of this study is to chart a way forward for the Project which will enable it to professionalise its activities. The data used in this approach can be regarded as both primary (the institutional planning documents which established the Project, the PTEEP tests, student records, the ESL-HG syllabuses and examination papers) and secondary (reports and research conducted on the primary documents).

In dealing with these data, both categorising and coding (content analysis), and 'rich narrative' (Smith, Connole, Speedy & Wiseman 1990) approaches are employed. The categories employed in the analysis of the syllabuses, ESL-HG examination papers and PTEEP tests are derived from the construct developed in Chapters Two to Six in respect of the PTEEP tests, and on the basis of the construct of communicative competence as articulated by the syllabus documents, in respect of the examination papers. The rich narrative approach attempts to place the events surrounding the establishment and continuation of the Project in context, using techniques such as ordering, synthesis, evaluation, and interpretation.

1.5.2 Quantitative Approaches

Chapters Eight to Twelve rely to a considerable extent on the use of statistical techniques to order and interpret the test performance data on which many of the validation procedures are based. The techniques used in this study are described below. In summary, the methods used in the various studies of performance are: correlation analysis (Pearson's product moment and

Spearman's rank correlation coefficient); analysis of variance and Kruskal-Wallis tests; discriminant function analysis and logistic regression; 2-sample non-parametric testing (Kolmogorov-Smirnov); chi-square analysis and the two-parameter Weibull model. These methods are described below in general terms, and more specifically in the contexts in which they were employed (i.e. in Chapters Eight to Twelve).

1.5.2.1 Correlation Analysis[6]

Correlation analysis enables the strength of the relationship between two variables - a pair of numerical values measured simultaneously on a set of individuals - to be described numerically as a correlation coefficient (a number between -1 and +1). This study uses two kinds of correlation approaches, the Pearson product moment correlation and the Spearman rank order correlation.

The Pearson product moment correlation, represented by 'r', is the most commonly used kind of correlation. It can only be used, however, under certain conditions. First, the two variables being investigated must be numerical and an interval or ratio scale. Generally, measurement variables assume any value along a stipulated continuum, such as height, or weight, or test scores[7]. By contrast, variables such as race or sex are categorical rather than numerical, and would not be appropriate variables on which to derive a Pearson product moment operation. In addition, the relationship between the two numerical variables must be at least approximately linear. That is, the pattern of the paired values should be able to be reflected by a straight line, showing that the higher (or lower) the value is of one variable - such as a score on Test A - the higher (or lower) the value will be on the other variable, such as Test B. In addition, linearity implies that an increment in Test A is associated with a corresponding increment (or decrement) in Test B, and the increment is identical.

The Pearson product moment procedure is used in this study to describe the relationship between scores on the PTEEP tests and the scores candidates obtain for their university courses or for previous language proficiency tests they had written. Many of the correlational analyses, however,

[6] General cautions and difficulties concerning the use of correlation analysis are detailed in Chapter Four.
[7] Technically, test scores are counts rather than measures, and if a zero score is viewed as indicating an absolute zero, might be treated as if they are on a ratio scale.

use the Spearman rank-order correlation (R). When the variables are measured on an ordinal (ranking) scale, and/or when sample sizes are small or the underlying data are not normally distributed, it is appropriate to use Spearman's R. In such a case the strength of the relationship between the ranks of two variables is measured. In this study, the data are ordinal in nature, being concerned with 'greater or less than' information. The use of Spearman's R is thus indicated. For example, the study is based on data that tell us that one student did better in the school-leaving examination than another as s/he obtained an A symbol and the other a C symbol. This ranking does not reveal, however, precisely how much better the first candidate was, just that s/he was better - that is, the scores are not based on an equal interval scale. Essentially, Spearman's R reflects the consistency with which an increment in Test A is associated with an increment in Test B even if that increment is not necessarily constant over the range of Test A.

In addition to these statistics, Kendall's tau is used to estimate the difference between two probabilities: that the probability that pairs of observed (bivariate) data (X,Y) exhibit the same order for the two variables, and the probability that the pairs exhibit reverse order; or that the observed data are in different orders for the two variables ($X_1 < X_2$ but $Y_1 > Y_2$). Kendall's tau was employed in the course of the "Survival Analysis" study discussed in Chapter Ten (section 10.4).

The correlation methods used to provide indices of the degree of reliability of the assessments apply some special procedures. In this study, the procedures used are:

- split-halves. This procedure allows reliability to be estimated from a single testing occasion. The test is split into two equivalent halves, and each student is assigned a score from each half of the test.
- the Spearman-Brown double-length formula. The split-halves procedure effectively reduces the length of the test by half, which tends to produce a lower coefficient. The Spearman-Brown double-length formula adjusts this lower coefficient by estimating the reliability coefficient of the whole test.
- Cronbach's alpha (coefficient alpha). This coefficient is used in preference to the Kuder-Richardson formulae 20 and 21, which are only suitable for use when a test is comprised

entirely on dichotomously scored items. As this is not the case with the PTEEP tests, coefficient alpha (Cronbach 1951) is used. Coefficient alpha is equal to the average of all the possible split-halves coefficients that could be generated for a particular test.

1.5.2.2 Analysis of Variance

Analysis of variance (ANOVA) tests are used when two or more means, or two or more groups, are compared. Basically, the tests compare the variability of scores within groups with the variability between groups. In the context of an admissions test, for example, one might be interested in investigating whether, if the admissions test scores were partitioned into quintiles, these categories (quintiles) would allow one to extrapolate about subsequent university performance. In other words, what ANOVA would test in this instance is whether the five groups of candidates (represented by the admissions test quintiles) differ significantly in their university performance.

ANOVA tests use the F-distribution to derive a value (the F-value) with which to test the null hypothesis. If the F value is 1 or less, it can be inferred that there is no evidence that the groups (or their means) differ and hence that they belong to the same population. In the example given above, if an F value of less than 1 was obtained, it would have to be concluded that the partitioning of admissions test scores into quintiles was not a meaningful operation for the objective of explaining university performance as measured by that index. That is, one could not extrapolate about university performance on the basis of quintile divisions. In contrast, F-values sufficiently greater than 1 indicate some evidence of groups being different.

The Kruskal-Wallis test (a non-parametric analogue to ANOVA) is based on ranks rather than means, and is used on data when the response variable (e.g. some measure of university performance) is ordinal or is not normally distributed. It is usually used to compare the overall distributions of values in two or more independent samples. For example, in this study, the Kruskal-Wallis ANOVA (H) was used to assess whether there were significant differences in

university performance between values of an index of 'ratio of courses passed to courses taken'[8], across the five levels (quintiles) of the PTEEP, treated as a categorical explanatory variable.

1.5.2.3 Discriminant Function Analysis and Logistic Regression

Discriminant function analysis and logistic regression are used in cases where one wishes to investigate which measurement variables discriminate between two or more specified categorical outcomes. In this study, the aim is to determine which explanatory measurement variables discriminate between two outcomes (success or failure at university). Success and failure can of course be variously defined, but in each definition of the response variable, students either pass or fail. The explanatory variables that are used in this study are the PTEEP scores and various measures derived from Senior Certificate scores.

1.5.2.4 Two-Sample Non-Parametric Tests for Differences of Distribution

Two-sample non-parametric testing examines the assumption that two samples were drawn from different populations. The Kolmogorov-Smirnov test is used to contrast two samples constituted by students (the top 20% of the PTEEP candidates and the bottom 80%), with respect to subsequent performance measures.

1.5.2.5 Chi-Square Analysis

Chi-square analysis is designed to measure association between variables with nominal data (i.e. data where there is a yes/no element - e.g. pass/fail, or top group/bottom group on the PTEEP). In analysing nominal variables, we are concerned with frequency issues - how many or how often - rather than with how much, as for measurement variables. In this study, chi-square analysis is used to test for significant differences in the frequencies of pass/fail responses within any one top/bottom tentative, explanatory grouping. In other words it tests whether the relative frequency of pass performances changes between the bottom and top PTEEP groups.

[8] This index is described in Chapter Ten (section 10.3).

1.5.2.6 *Survival Analysis Approaches*

Survival analysis is applied to an estimation of the length of time it takes for students to be excluded from university: that is, on the calculation of time-to-exclusion and the relationship of this to specified variables (e.g. admissions test scores). The data are observed at successive points in time until the 'event' i.e. exclusion from the institution, is achieved. Special statistical techniques are required for the analysis of any model for survival information data (that is, to model the exclusion process). The model selected and employed by Polakow (1999) is the two-parameter Weibull model. This approach and findings are described in Chapter Ten (section 10.4) in the investigation of predictive validity.

In this chapter, the reasons for and aims of the study have been outlined, as have the methodological approaches adopted. Chapter Two sets out to lay the foundation for the process of building the construct of an academic literacy test – it approaches this by exploring several important issues in the area of selection for Higher Education.

CHAPTER TWO

GENERAL ISSUES SURROUNDING SELECTION FOR HIGHER EDUCATION

2.1 Introduction

Chapter Two contains an analysis of various issues related to selection to Higher Education. Specifically, it questions whether selection is necessary and what options exist in this regard, and discusses the factors that must be taken into consideration in order to satisfy demands of fairness and equity while meeting the needs of society for skilled and well-educated citizens.

2.2 The Purposes of Higher Education

Using Strike's (1983, cited in Coombs 1994) categorisation of the fundamental purposes of higher education as a basis for considering issues of equality, Coombs (1994) proposes the following three main categories of purpose:

- preparing students for responsible citizenship in a democratic society. To this, Coombs adds preparing them to be good friends, colleagues, and family members;
- nurturing and developing students' abilities so that they can contribute to the economy and culture of their society; and
- inducting students into the intellectual and cultural traditions of their society so that they may lead more satisfying and enjoyable lives.

These three purposes, however, do not capture an important function of higher education, which more than any of those above, distinguishes it from schooling. This distinction is included in the White Paper on the transformation of higher education (DoE 1997b:1.3), which adds the following essential objective of higher education – that it should "...contribute to the creation, sharing and evaluation of knowledge ... through research, learning and teaching". Morrow (1994:43) makes the point as follows: "The founding rationale [of universities] is not to supply the economy with person power, to shore up or undermine a particular political dispensation ..... . The guiding ideal of universities is to *constitute* the realm of academic learning; to provide an institutional home for academic practices and access to them." (italics in original).

In a similar vein, but with specific reference to the needs of the African context, Ajayi, Goma and Johnson (1996:173) state that

> "... the efficiency of universities is no longer to be measured by the number of graduates produced, but by their quality and capacity to produce the knowledge required to reduce the widespread dependence and marginalisation of the African countries and continent."

In emphasising this function of universities in Africa, Ajayi et al warn of the tensions it poses for selection. They argue that maximising opportunities for the greatest number of people to

participate in Higher Education - a crucial part of any country's 'status mobility system' (Herman 1995) - on the grounds of such matters as equity and social justice might militate against quality and the generation of new knowledge that the societies so desperately need.

2.3 Objectives of Selection

All institutions naturally strive to select the 'best' candidates from their pool of applicants. Determining which candidates are in fact the best, however, is no simple matter. As is discussed below, the criteria for selection have to bear many, often competing, factors in mind. Nevertheless, a starting point for most enrolment management operations is their institution's Mission Statement. In the case of UCT, the overarching mission is " ... educating for life and addressing the challenges facing [South African] society" (University of Cape Town Mission Statement). While this objective is not particularly helpful in concrete terms, the full text of the Mission Statement makes it clear that both diversity and academic merit, in terms of both developed ability and potential, are primary aims, as is a commitment to redress of past discrimination and to undertaking an active developmental role in rebuilding South African society.

Translating these aims into criteria for admissions and selection is a complex and difficult task. Klitgaard (1985:183) points out that "... the only objective of selection that we seem able to predict is academic performance". However, even predicting academic performance is very imperfect, as is discussed in 2.6 below. What seem impervious to prediction, however, are such outcomes as success in later life, or contributions to society. Indeed, the finding of the study by Bowen and Bok (1998), that African Americans reported a greater level of involvement, post graduation, in social issues and thus that there are substantial social rewards from applying affirmative action in admission to selective institutions, is challengeable on the grounds that a similarly greater proportion of African Americans in their study had taken degrees in Humanities areas. Thus, their greater involvement might reflect their professional training and interests as much as their purportedly greater social commitment (W. Gevers, personal communication, December 1999).

At least partly as a result of the complexities involved in translating institutional goals into admissions criteria, and the consequent difficulties in justifying some procedures, selection to Higher Education is a controversial and contested terrain. Because of the high stakes attached to admissions decisions, as well as the desirability and importance of having clearly defensible procedures, it is useful to consider some of the alternatives to selection before embarking on the complex process of developing selection criteria.

2.4 Alternatives to Selection

Selection occurs whenever choices are made – and choices are made whenever the number of applicants exceeds the number of places available. Even in seemingly open institutions - open in the sense of taking all applicants who meet a minimum requirement such as the exemption in the South African context - selection takes place for high demand courses. It is preferable that the basis for these choices is clearly articulated, although this is not always possible. In South Africa, it is a legal requirement (Higher Education Act 101 of 1997) that Higher Education institutions must make public their admissions policies – and this requirement includes the selection criteria on which these policies are based.

South Africa's history of apartheid has made its citizens acutely aware of discriminatory practices, and selection for Higher Education is an area that has long been controversial. South Africa is not alone in this area of controversy, of course – in the USA the under-representation of minority groups particularly, but not only, in selective institutions is a recurrent issue. Higher Education is a major pathway to upward social mobility. As a key component of what Ogbu (1982) and others call the 'status mobility system' of a society, it is a crucial tool in the rapid normalisation of South African society.

As was suggested above, the necessity for some form of selection can most clearly be demonstrated by investigating the alternatives. Klitgaard (1985) suggests the following possible alternatives to selection: raising tuition prices, growth, random selection, and tryouts. A fifth alternative, 'bureaucratic inefficiency" (Segall 1994), could be added to these alternatives.

2.4.1 Raising Tuition Prices

One way to avoid the need to select would be to set fees at a level that could only be met by the target number of applicants. It could be argued that particularly in a poor country, the levying of any fees, no matter how modest, constitutes a mechanism for restricting numbers. However, even wealthy countries such as Australia and the UK, with comparatively generous budgets allocated to Higher Education, have found it necessary to introduce fees, despite widespread student opposition. The provision of an adequate financial aid system for genuinely needy students is commonly used to safeguard against *de facto* selection through the levying of fees. Even when this safeguard is in place, however, the debt burden of poor students[9] can be a serious obstacle to good academic performance as it adds to the anxiety and stress of such students.

As a deliberate means of avoiding selection, however, raising tuition prices is an unacceptable response and can create more problems than it solves. For example, it is almost certain to have the effect of perpetuating privilege, or of reducing diversity in the student body.

Nevertheless, ability to pay does play an important role in admissions, both in South Africa and elsewhere. An example of this in one Higher Education institution in South Africa was described at a 1994 workshop on admissions issues (Barsby, Haeck & Yeld, 1994)[10]. Segall (1994) reports a similar, more general situation in an overview paper on admissions practices in South Africa.

Few people would recommend, as a policy, the admission of students on the grounds of their ability to pay or through the raising of fees – and particularly not when it would fly in the face of the Higher Education Act with its emphasis on redress. Raising fees is thus not a viable alternative to selection.

[9] Financial aid usually consists of at best a combination of loan and bursary, at worst a loan only.

[10] The Dean of the Faculty of Engineering made offers to all the candidates he had selected from the applicant pool for his Faculty. On registration day, however, only about half of the candidates actually registered. In order to fill his places, the Dean then stood up in the hall where registration was taking place, and asked for all applicants who had originally applied for Engineering and had not been offered a place, but who were present and who had the required registration fee, to come forward and register. In this way he filled his first year quota. When questioned about why these applicants, who had applied for Engineering but had initially been rejected, were there in the hall, the admissions officer replied that this - last-minute offers conditional upon registration fee - was a well-known practice, and so students assumed it would take place, and came prepared. The end result, of course, was that many students with weaker academic claims were accepted because they had money and were in the right place at the right time.

2.4.2 Changing the Size of the Higher Education System

Expansion and contraction of Higher Education systems occurs for various reasons. When there is a large pool of well-prepared candidates seeking entry, and when the economy is able to absorb at least a large proportion of its graduates, the system can absorb large growth without too much harm being done. Existence of such a pool is seldom the case, however, and the deteriorating quality and confusion that has beset the Higher Education systems of many African countries bears witness to the harm caused by growth for political reasons, and in the absence of enabling conditions such as a growing economy. As Ajayi et al (1996:204) state, "The coincidence of deteriorating quality and accelerating demand is ... a most serious threat to higher education in Africa ...".

In recent years, the Higher Education sector in South Africa has been experiencing a largely unanticipated drop in the numbers of applicants (and registered students). This drop is particularly acute at the Historically Black Universities (HBUs), but is not confined to them. Robbins (1999) states that the HBUs had an average decline in student numbers of 11% for 1999 alone - in 1998 numbers had declined as well. The Historically White Universities (HWUs), in contrast, maintained their numbers, with small but steady growth occurring at the previously Afrikaans-medium institutions.

One of the reasons suggested for the falling numbers of students at HBUs is the issue of fees[11]. In the past, many HBUs in particular allowed students to accumulate large fee debts, and looked to the new national Department of Education for assistance. To date, however, the Ministry of Education has shown little sympathy for the plight of these institutions – on the contrary, in a recent speech the then Minister, Professor Bengu, stated "I must express my profound disquiet at the financial status of some of our institutions. ..... Too many institutions are running unacceptably high overdrafts" (Bengu 1999, cited in Robbins 1999). The financial scandals at some institutions, as well as evidence of serious mismanagement, has made the Ministry unwilling to simply bale out the

[11] Other reasons include a decline in the number of school-leavers who qualify for university admission, the proliferation of private higher education providers, and a high incidence of unemployment, even for graduates.

institutions, which are therefore forced to extract the fees from students instead of extending sympathetic long-term loans.

In relation to selection, the phenomenon of falling demand has reduced the pressure to select at the HBUs. Recent evidence indicates that some institutions are now deliberately admitting students with very little chance of success in an academic environment of reasonable quality, in order to boost their student numbers and thus their state funding (Yeld 1999).

Fluctuations in demand make the development of criteria for selection very difficult. When there is increased demand, institutions have two main options: they can grow to avoid or minimise choice, or they can avoid growth by developing more effective selection procedures and criteria to cope with the increased demand and continue to control access. When there is reduced demand, there are also two main options. First, the decision can be to fill places irrespective of quality, thereby avoiding selection. Second, it could be decided to develop more effective selection options that will identify academic talent in a probably weaker pool in terms of developed ability (i.e. to avoid a drop in student numbers by establishing access routes further down in developed ability).

Both of these scenarios (growth or reduction), however, can only postpone, not remove, the need for selection. If demand is too small, going ever downwards in the applicant pool would only be possible if very well developed methods of identifying academic potential were available as a basis for selection, and of course if appropriate courses to meet the needs of the students had been developed. If demand is great, on the other hand, there will come a point at which it is not possible simply to accommodate all, and selection will become necessary.

Using size as a mechanism to avoid selection is thus only a very limited and short-term alternative to selection.

2.4.3 Random Selection

Many variations of random selection are possible, from genuinely random for all applicants, to the version adopted in the Netherlands, where a weighted lottery system is used. The rationale for

weighting is usually that certain courses of study are not best allocated on a wholly random basis. For example, medical schools have long held that academic merit is only one of the attributes that make a successful doctor. In the Netherlands, the lottery system is only used for certain oversubscribed systems, and the system weights on the basis of end-of-school results.

In South Africa, a similar system was tried at the University of the Western Cape (UWC), where a lottery was introduced for relatively low demand courses: the lottery did not extend to the Faculty of Dentistry, for example, and only in a very limited way to Science. In this case, 20% of the first-year students were selected on the basis of their Senior Certificate results, gender, population group classification, geographical origin, rural/urban origin, and social class. The remaining 80% were selected on the basis of a lottery (Herman 1995). Interestingly, Herman (op cit:267) states that "... no applicants with high (school) pass marks .... were excluded". If, therefore, UWC's applicant pool had included a higher proportion – for example 30% - of candidates with high pass marks, this particular allocation of places would not have been feasible, and the proportion allocated to lottery distribution would have had to be reduced. Since all applicants with good school results were offered places, and since only relatively low demand courses were included in the lottery system, the implication is that the lottery was, primarily, a pragmatic rather than a principled response. That is to say, it operated only at the level of low achievers, whose school results are very poor predictors anyway of later success (Yeld & Haeck 1997, Herman 1995, Barsby et al 1994, Badsha, Blake & Brock-Utne 1986).

The lottery system is intuitively appealing, particularly in its weighted form: it suggests, simply, that one's chances of being admitted do not depend on one's opportunities prior to application, except at the top extremes. However, the lack of connection between progress and effort at school and later life chances can have extremely serious consequences. This issue is discussed further in section 2.5.2 below.

2.4.4 Fail-first Approaches

In this approach, selection is in effect delayed for a period of time: however, it appears on the surface to be an alternative to selection. Essentially, a 'tryout' approach admits applicants either

by random selection or by some other means, and then assesses their performance at some later time (e.g. at the end of first year), on the basis of which the candidate is accepted or rejected. The Teach-Test-Teach (TTT) project at the University of Natal, South Africa, was in its early years a version of the tryout approach (Griesel 1991, 1999, Zaaiman 1998). In this model, a large group of students was tested, put through a fortnight of instruction, then re-tested. Selection was planned to take place after this two-week tryout period. In fact, in the TTT project, very few students were rejected at this stage. This was understandable, as in a research project of this kind, the performance of 'rejected' students is as important as that of selected students in terms of assessing the value of the selection model.

In the real situation, however, delaying the time of selection will not lessen the impact of rejection – in fact, given financial realities, it might make it more difficult (for example, the students will have accumulated fee debt). The fundamental point is that the use of expensive Higher Education resources in the form of places in programmes is an inefficient approach to selection, and effort would be better spent in improving other instruments such as the Senior Certificate, and/or in developing additional instruments. Reducing the costs of selection by increasing the costs of instruction, for example to the state, the student, and the institution, by teaching numbers of students with little chance of success, does not make sound economic sense. Not only does it contribute to the "... dilution of instructional and other resources ..." (Herman 1995:268), it contributes to the construction of a first-year class that is extremely difficult to teach, with an arguably high proportion of students with no chance of success. Thus, while acting as a slow release admissions filter, the 'tryout' method might in fact result in a decreased quality of learning and teaching for more able students, as well as a costly and demoralising experience for their weaker peers.

2.4.5 Bureaucratic Inefficiencies

In addition to the four approaches to avoiding selection suggested by Klitgaard, that of selection by default – or by system or administrative inefficiency – could be added. Segall (1994) gives examples in this regard of institutions failing to respond to applicants, who then arrive at

registration and are admitted, sometimes on a first-come-first-served basis, sometimes on production of the required fees. Many other applicants, however, are discouraged when they do not hear from the institution, and thus are effectively de-selected through bureaucratic inefficiency. Clearly, this option may avoid selection, but it does not represent a sensible approach to planning.

The discussion above on alternatives to selection illustrates the inevitability of some form of selection in situations where the number of applicants exceeds the number of places, and/or where necessary skills and/or knowledge are required.

2.5 Factors in the Development of Admissions and Selection Criteria

Once it has been concluded, however reluctantly, that there are no viable options to selection as a means of allocating highly sought after places in Higher Education, many decisions remain to be made in terms of the criteria on which selection will be based. In considering options in this regard, thought must be given to the impact and consequences such criteria might have on applicants and the broader society; the composition of the student body; the contribution Higher Education can make to national goals; and the ways in which academic merit is conceptualised.

2.5.1 The Public School System, and Exit Levels

It is a truism that the level at which tertiary education can begin is inseparable from the level at which secondary education ends. Detailed and current knowledge of the school system is an essential tool for the development of selection criteria. This knowledge should include keeping abreast of likely future developments as well as developing the capacity to conduct studies of performance trends.

2.5.2 Consequences for Applicants and Society

The introduction of criteria for admissions can have unexpected consequences. An example in this regard from the United States is the National Collegiate Athletic Association's decision (Proposition 48) to raise the entrance requirements for aspiring college athletes. In contrast to the fears of many that this would "... reduce minority college opportunities ..." (Heubert & Hauser 1999:20), a study by Klein and Bell (1995, cited in Heubert & Hauser 1999) found that not only did the higher

admission requirements not reduce the numbers of minority athletes gaining admission, it appears to have led to higher graduation rates amongst this group. Reasons for this change are apparently that the potential athletes took more challenging courses in school to meet the higher admissions standards, and this choice helped to prepare them more effectively for tertiary study demands. Thus, by changing admissions requirements, the high school careers and experiences of potential applicants were enhanced.

In the South African context, an example of how decisions related to admissions criteria could impact on applicants and the broader society, would be a decision by the Higher Education system, or some subset of it, not to include Senior Certificate (SC) results among its criteria for admissions. If this decision were to be taken, it is likely that the SC would suffer greatly – and with it, the secondary school system[12]. While criticisms of the SC are well founded, the value of the existence of "... an end examination with clout" (Yeld & van Bommel 1997) should not be underestimated. There are many reasons for this claim, chief among them being:

- the motivational force exerted by the examination;
- the potential of the examination to support educational innovation; and
- the role of the SC in South Africa's fragile quality assurance system. As Taylor (1999:199) states: "In terms of learning outcomes, there is at present only one quality assurance mechanism in place in the entire system – the matriculation examination at the end of Grade 12".

The point being made here is simply that admissions criteria which do not include SC results are likely to undermine the SC itself in the eyes of learners and the general public, and this consequence is highly undesirable, particularly in a time of great educational change. As Morrow (1994:36) notes:

[12] A less radical example of an admissions criterion that would nonetheless impact on candidate preparation would be a decision not to recognise any SC Standard Grade (SG) subjects for admissions purposes. At present SC candidates take a prescribed mixture of SG and Higher Grade – HG – subjects for endorsement purposes. It is not difficult to imagine a sudden change in the numbers of candidates registered for various HG and SG subjects. One immediate short-term consequence would be increased failure rates, as many teachers are not qualified to teach at HG level. In the longer term, however, the impact of the admissions criterion might be positive, as the crisis this would cause might act as a catalyst for teacher education and in-service training.

> "... a university access policy which is sceptical about the value of achieving a matriculation certificate (even when there are good grounds for such scepticism) can seriously undermine the aspirations and ambitions of both teachers and learners in the schooling system (and thus ... contribute to the decay of a culture of learning in schools)."

The impact of selection criteria on the Higher Education system as a whole needs to be taken into account as well. If - as is usually the case – high demand institutions establish criteria that identify the top school-leavers, and if they manage to attract these students, it means that the remaining institutions have no choice, if they want to fill their first-year places, but to develop less stringent criteria. The 'pecking order' that results from such a situation can be mitigated through the establishment of niche expertises, where an institution becomes known for the excellence of its Agricultural Faculty, for example, or for its small classes and individual attention in a rural setting. In South Africa, however, the gap between the HBUs and HWUs, created deliberately under the apartheid regime, appears to be widening. Having, as Zaaiman (1998:11) suggests, "... shoulder[ed] the moral burden of accepting the often under-prepared school leavers", the HBUs are now regarded as "... being a manifestation of apartheid inferiority", and students are going where "... they can get the best education for their money" (Robbins 1999). One upshot of this strategy is that "... higher achieving students of both the middle and working class prefer enrolment at HWUs in South Africa in lucrative professional fields" (Herman 1995:268).

This preference is illustrated in the following table, which shows that the historically white universities and technikons (HWTs) now enrol a greater proportion of the country's African students than do historically black universities and technikons (HBTs), in strong contrast to the position as recently as 1993.

Year	Numbers of African students in Higher Education	Proportion at HWUs	Proportion at HBUs	Proportion at HWTs	Proportion at HBTs	Proportion at distance institutions
1993	147,000	8%	41%	5%	8%	38%
1999	200,000	24%	21%	15%	12%	28%

Table 2.1: Numbers of African Students in Residential Universities and Technikons, 1993 and 1999 (adapted from Naidu 2000)

The examples above illustrate the need for consideration to be given to the social consequences of various admissions decisions. In South Africa, certain directions have been laid down by national

policy, such as the need for redress, representivity, and equal access, but institutions nevertheless have considerable autonomy.

2.5.3 Impact on the Composition of the Student Body

It seems uncontroversial that public educational institutions should aim to constitute their student bodies in such a way as broadly to reflect the general population. In South Africa, the Preamble to the Higher Education Act (101 of 1997) states that one of the aims of the Act is to "redress past discrimination and ensure representivity and equal access". Apart from considerations of social justice and individual human rights, however, many studies have been conducted to investigate the impact of diversity on students' educational experiences, and thus to test the assumption that positive outcomes will result from mixing students from different racial groups.

In brief, the studies that have been conducted to date related to diversity can be grouped as follows:

- Studies which support diversity on the grounds that it enriches students by exposing them to different life experiences, thereby challenging them intellectually and facilitating the development of mutual respect (e.g. Chang 1999, Tierney 1997, Moses 1994);
- Studies which argue that diversity contributes to a dilution and distortion of academic standards, on the grounds that black students are more likely to be educationally disadvantaged, or to have been admitted with lower scores (e.g. Sowell 1989), and that many so-called integrated campuses are in fact racially polarised. This polarisation, it is argued, tends to confirm stereotypes (Steele 1990, Thernstrom & Thernstrom 1997) rather than to facilitate positive experiences.

However, what both camps would agree on is that simply mixing students from different racial or cultural groups is not guaranteed to produce educational benefits. On the contrary, research strongly suggests that "...when efforts to improve diversity are taken seriously and done well ..." (Chang 1999:379), the educational environment for all students will yield benefits.

Returning to the development of selection criteria that will facilitate and build diversity in the student body, it is clear that reliance on achievement test results alone will result in low numbers of educationally disadvantaged applicants being selected for admissions. This consequence has been repeatedly demonstrated. Bowen and Bok (1998), in their study of highly selective institutions in the United States, found that if all the institutions adopted 'race-blind' admissions policies, the proportion of Black students admitted would drop from 7% to 3%. Johnson (1998) reports dramatic figures from a study at the University of California, Berkeley, where as a result of the ban on the inclusion of race as a factor in admissions policies, there was a 64% drop in the number of Black students admitted in 1998 compared with 1997 (191 compared to 562). At UCT, some 2,000 ex-DET students, over the last nine years, would not have gained admission to the institution if school-leaving results were the only criteria used (Yeld & Visser 1999).

If diversity in the student body is valued, however, it is necessary for some other information, or nuanced use of test scores, to be developed. A decision to use test scores as they stand will have known impacts on the composition of the student body in that they will reduce the representation of educationally disadvantaged candidates.

This reduction is illustrated in Table 2.2 below, which shows the PTEEP performance of different groups of applicants to UCT for the 1998 academic year. The scores are reported as deciles[13]. In the table below, in the 'Overall' column, there were 1,311 candidates – therefore there were 131 candidates in each decile. Decile 1 reports the score of the top candidate (93.1%) as the top of that decile, and the score of the candidate who came 131st (77.9%) as the bottom of Decile 1.

[13] Deciles are obtained by ranking all the scores of a candidate pool, dividing the number of candidates by 10, and then reporting the scores of the candidates at each tenth interval.

DECILE	OVERALL (%)	EX-DET (%)	EX-HoA[1] (%)
1	77.9 – 93.1	63.3 – 81.3	80.8 – 93.1
2	70.9 – 77.3	56.8 – 62.6	75.2 – 80.0
3	64.5 – 70.3	51.8 – 56.1	71.2 – 74.4
4	57.6 – 64	47.5 – 51.7	68.0 – 70.4
5	51.2 – 57	42.5 – 46.8	64.8 – 67.2
6	44.2 – 50.6	38.3 – 41.7	60.0 – 64.0
7	36.6 – 43.6	33.8 – 37.4	54.4 – 59.2
8	32.1 – 36	28.1 – 33.1	48.0 – 53.6
9	23.3 – 29.7	23.2 – 27.3	40.8 – 47.2
10	4.1 – 22.7	8.6 – 22.3	40.0 – 6.4
No. of candidates	1311	610	414

Table 2.2: Performance of Candidates on the PTEEP Test: 1998 Entry

Note:

[1] *– ex-HoA refers to students who attended schools which would, prior to 1996, have been under the authority of the House of Assembly (historically white schools).*

Dramatically different levels of performance are clearly illustrated in this table. If the institution had decided to make offers only to the top two deciles of the overall group – i.e. to those candidates obtaining 70.9% or more – almost no ex-DET students would have been selected. However, the selection criteria agreed by the institution stipulate the top two deciles of each column, i.e. of each candidate pool. So ex-DET students with 56.8% and over would have been considered, and ex-HoA with 75.2% and over. It is important to note that these categories are not based on racial classifications, but on the education authorities to which the schools they attended would have been allocated prior to 1994. Thus an African candidate attending an ex-HoA school (an historically 'white' or private school) would be placed in the ex-HoA rank, whereas an African candidate attending an historically DET school would be placed in the ex-DET rank.

One of the problems with this approach is, of course, that candidates obtaining 56.8% do not have the same level of developed abilities as those obtaining 75.2%. This, however, is a placement, rather than an admissions, issue. In the example above, applicants with relatively low scores but high decile rankings, are likely to be offered a place in a foundation or extended programme of some kind. For this reason, two scores are reported for all candidates: one, their decile ranking in the overall testing group, and the other, their decile ranking in their educational category[14].

[14] By way of illustration, Sipho Bloggs, who obtained 58.1% on the test, and who attended an ex-DET school, would be reported as having an overall decile ranking of 5, and an ex-DET ranking of 2. This pair of results would result in his being made an offer of admission, but being placed onto a foundation programme.

In addition to diversity, an essential characteristic of the student body that must be considered in the development of selection criteria is that of academic ability. There can be no denying that the academic quality of the student body is a crucial element of the teaching and learning environment. As Adelman (1999:B5) comments in the context of the USA, "... the haves in our society are divided from the have-nots by virtue of the SAT scores of their college companions ..." as much as by virtue of their own SAT (the SAT1: Reasoning Test) scores. Setting minimum academic criteria for admissions is thus no light matter for selective institutions: they need to maintain quality while achieving their desired mix.

The tensions between these two imperatives (diversity and academic merit) will continue to exist for as long as opportunities tend to be arranged along racial and class lines. Managing the tension calls for creative responses, and for ways, such as that described above, of identifying talented students while not denying the effects of educational disadvantage.

2.5.4 Contribution to National Goals

Higher Education in South Africa has much in common with other countries in terms of its purposes. What is somewhat different is the clearly articulated commitment to redress. This commitment is emphasised in the Constitution, which in its Bill of Rights provides for legislative and other measures which will ensure the protection or advancement of previously disadvantaged persons – disadvantaged, that is, by unfair discrimination. In addition, Higher Education institutions, while permitted to determine entry requirements beyond the statutory minimum, are required to "...ensure that selection criteria are sensitive to the educational backgrounds of potential learners ..." (DoE 1997b:29).

South Africa has as yet no formally differentiated system of Higher Education, where individual institutions would aim to make different but complementary contributions to national goals. In broad terms, however, selection criteria would seek to ensure representivity, to promote growth and development in the areas of Mathematics and Science, and to meet the urgent development needs of the country.

2.6 Identifying Predictors of Academic Performance

Few people would disagree with the proposition that academic merit – or the likelihood of success - should be taken into serious consideration in selection to higher education. Defining and identifying academic merit, however, are far from simple matters, and create much controversy. Following sections in this study are devoted to considering what predictors of academic merit or achievement are available, and the points made here are not intended to be comprehensive or thoroughly argued until later.

In South Africa, as in many countries, the school-leaving examination has traditionally been used for a dual purpose: to certify a level of achievement on the course of study preceding the examination, and to act as a gatekeeper for post-secondary educational opportunities. Where there is clear evidence of grossly unequal provision of educational opportunities, however, the latter purpose, in particular, is problematic. In this context, institutions may respond in various ways. They may, for example, seek other indicators of achievement, such as the SAT tests in the United States. Additionally or alternatively, they may decide to develop ways of assessing candidates' potential to succeed, as opposed to relying solely on developed ability. A further option would be to assess the personality or character of an applicant, as is done in admission to many medical schools in South Africa and elsewhere, on the grounds that personal disposition greatly affects an individual's chances of success in a challenging environment.

In reality, few admissions systems rely on only one source in attempting to assess the academic merit of candidates. As is argued in later chapters, the greater the disparities in educational provision prior to selection, the more necessary it is for a system to develop and introduce a variety of methods of selection to Higher Education.

2.7 Selection and Fairness

2.7.1 Adverse Impact, Discrimination and Bias

Adverse impact refers to a situation where a smaller proportion (of the applicants) of one group is selected than of another group (see, for example, Bartram 1995). By way of illustration, say that

3,000 applicants compete for admission to a medical school, and that 2,000 of these are men, and 1,000 women. However, 1,600 of the men are selected (80% of the male applicant pool), and only 200 of the women (20%). The difference in these selection rates translates to a factor of 0.25 (from the ratio 20% : 80%), or 25%. Adverse impact is said to have occurred when the selection rate for one group is less than 4/5ths (80%) of another group's selection rate. In the hypothetical case of the medical school illustrated above, women have suffered from adverse impact.

However, if the applicant pool of one group is not comparable to the applicant pool of another, it is not correct to assume that differential selection rates are caused by (or reveal) adverse impact. Locating the above example in the South African context illustrates this form of false inference. For a number of reasons (such as an almost total lack of vocational guidance in schools, and a scarcity of relevant role models), a very large proportion of black applicants to Higher Education apply to study medicine as their first choice. In this case, even if the great majority of places in the MBChB programme had been allocated to black applicants, the acceptance rate for this group might still fall foul of the 4/5ths rule. Clearly, some nuancing of the method would need to be applied in this situation – not to mention the need for urgent attention to the provision of appropriate school guidance.

Nevertheless, the adverse impact ratio is a useful indicator, and could be routinely used as a check on the selection process. If problematic ratios are found, the reasons would need careful investigation[15].

The gender example above also illustrates the concepts of discrimination and 'unfair discrimination'. If the selection rate for women was lower because a high proportion of women applicants were not eligible for selection (for example they had not taken Mathematics at the Higher Grade at school), the lower rate could not be ascribed to discrimination. If, however, it can

[15] In the hypothetical example given above, the group differences were gender based – given South Africa's history, racial classification will tend to be the over-arching group marker, but other factors should in fact include those required by law. For example, the Constitution of South Africa recognises the following: race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual orientation, age, disability, religion, conscience, belief, culture, language and birth (Constitution of the Republic of South Africa 108 of 1996). Until the country reaches a more stable stage, however, it is unlikely that scarce resources will be found to give sufficient attention to all these categories of potential discrimination.

shown that girls' schools have systematically not offered Mathematics at the Higher Grade, the fact that female applicants had not taken it and were thus not eligible is the result of unfair discrimination – and thus they are being unfairly discriminated against in not being selected. In this case the onus would be on the institution to provide opportunities for females to 'catch up' in this regard.

It can thus be seen that failure to recognise past educational opportunities can result in unfair discrimination. A case in point here is the use of standardised achievement tests when it is known that many candidates will not have covered the curricula on which the tests are based. This use occurs in an extreme form in South Africa at present, where a large proportion of the candidates has not been prepared adequately for the Senior Certificate examination. The report of the Ministerial Committee on the Senior Certificate Examination (DoE 1998) makes chilling reading in this regard. Listed below are some of the reasons given by the Committee for the poor performance of many candidates on the examination:

- Lack of appropriate standards of assessment prior to Grade 12.
- Inadequate coverage of the syllabus. In Mathematics, the committee found that many schools had still not started their Geometry syllabus even though the examination was 20 teaching days away.
- Lack of textbooks. Approximately 60% of South African schoolchildren in 1999 did not receive textbooks. The situation is somewhat better in Grade 12, but the candidates in a given year are unlikely to have had textbooks in previous years. It needs to be emphasised that the importance of textbooks is particularly acute in a context of large classes and poorly trained and/or frequently absent teachers.
- The very low number of teaching days in many schools. Taylor and Vinjevold (1999:136) describe the results of a workshop during which school principals examined the timetables of their schools for the preceding year. It was found that of the 191 days on which teaching could have taken place, only 21 (on average) were fully devoted to instruction. The rest were taken up by such events as examination writing, teacher strikes, slow starts to terms, examination

preparation, teacher marking time, and pay-days. Clearly, candidates who have only had 21 out of a possible 191 days to prepare cannot fairly be expected to have covered the same learning ground as their more privileged peers.

It is very difficult not to argue that a selection system based entirely on the results of the SC examination must be open to a charge of unfair discrimination. However, dispensing with the examination because it reveals inequities in schooling is akin to killing the messenger. Rather than undermining the SC, or abolishing it, this study sets out to recommend additional measures which can be used, in conjunction with the present SC or its proposed successor, the FETC, to ensure that talented yet disadvantaged candidates are not further disadvantaged in the selection process.

2.7.2 Equal Opportunity and Affirmative Action Models

There are several different models of equal opportunity (cf. Coombs 1994, Stowell1992, Zaaiman 1998). The most conservative of these models considers an individual to be able to demonstrate ability and merit irrespective of prior educational opportunities. This is called the 'qualified individualism' view of equal opportunity. In this view, selection based on demonstrated school achievement is perfectly acceptable, as everyone is considered to be being treated equally (e.g. they are all writing the same Biology HG paper) and thus to have an equal opportunity to succeed in the competition for places. As was shown in Table 2.2 above, selection on this basis would have resulted in an almost entirely white first-year student body at UCT in 1998.

A more progressive view takes previous educational opportunities into account, and recognises that different groups may be differently prepared and thus have unequal opportunities to succeed in a competition for scarce places. In this situation, it is recognised that unless different (alternative or additional) criteria (predictors) are employed in the selection process, the process cannot be considered fair. Candidates in the SC examination who have attended schools such as those described above, for example, would be recognised as being educationally disadvantaged in relation to candidates who have attended less dysfunctional schools. This position of 'unqualified individualism' – where fewer qualifications or restrictions are placed on the choice of predictors or criteria for selection – enables admissions officers and policy makers to build in compensatory

devices (such as what are termed, in the United States, 'race-sensitive' criteria). At UCT, this takes place in a number of ways, one example of which has been discussed above (section 2.5.3).

Both the positions of qualified and unqualified individualism can be classed as 'input' models – what goes into the development of the criteria which guide the selection process. However, at the more radical end of the spectrum is the view of equal opportunity that focuses on outcomes – in this case the outcomes of the selection process. The focus is on providing access to opportunities in an equitable way, rather than providing an effective selection mechanism. So, for example, it is more important to ensure that the composition of the student body reflects the composition of the wider population in terms of some agreed on characteristic such as racial classification, than to select a student body which will graduate in a reasonable time.

There are many problems with this. As Zaaiman (1998:181) notes, "[S]election should be seen as a contract to teach at that student's level". Constituting a first-year class according to some criterion unrelated to academic ability would seriously endanger the fulfilment of this contract.

In addition, a critical issue in relation to allocating places on race-sensitive grounds is that applicants with better qualifications would be rejected. In contexts where educationally disadvantaged groups form the minority, as is the case in the UK, the United States and Australia, the number of such occurrences will be small. The Bowen and Bok study, for example, found that if all the institutions adopted 'race-blind' admissions policies, the proportion of Black students admitted would drop from 7% to 3%. Hacker (1998), using the Bowen and Bok data, estimated that as only approximately 700 out of nearly 30,000 places were allocated on an affirmative action basis, only the same number of white or Asian students would have been affected. In other words, in order to maintain black participation at the 7% level, it is necessary only to disadvantage relatively few whites or Asians (the participation rates of these two groups would drop by approximately 2.5% i.e. from about 93% to 90.5%).

However, in societies where the majority has been educationally disadvantaged, achieving representivity (or dramatically increasing the participation rate of disadvantaged applicants) while

educational provision remains unequal, will almost certainly result in the admission of applicants with little chance of success, and serious deterioration of the teaching and learning environment. Until secondary schooling improves dramatically, and thereby the numbers of adequately prepared applicants, therefore, race-sensitive admissions criteria can best be used as part of the input side of the selection process.

2.8 Educational Disadvantage and Selection

As Hofmeyr and Spence point out (1989), views on what is now known as educational disadvantage have developed from the deficit-based views prevalent in the 1980s, when students were characterised as cognitively deficient in some way. The term 'educational disadvantage' in South Africa, while still sensitive, is now generally accepted to refer to the outcomes of the long-term under-resourcing, mismanagement, and deliberate oppression of the system designed in the years of apartheid to cater for the educational needs and aspirations of black learners. This change in understanding represents a shift from attributing the cause of the disadvantage to some lack of capacity within the individual, to attributing the cause of the disadvantage to the environment in which the individual, willy-nilly, underwent her/his educational experiences. As Ndebele (1995) states, no use of the term educational disadvantage should take place without a clear understanding and highlighting of the reasons for the disadvantage. Failure to specify that understanding is likely, he suggests, to lead to a simple conflation of black with disadvantage, and to create or consolidate a view of black people as being inevitably disadvantaged.

The causes of educational disadvantage are many, and have been extensively documented internationally as well as locally. They include the following:

2.8.1 Socio-economic Status (SES)

Indices of SES are usually based on the educational, occupational and economic status of the parents or responsible adults in a family. In the National Adult Literacy Survey (NALS) conducted in the United States, it was found that irrespective of the educational level of the respondents, those with 'better educated parents' scored higher on prose literacy tasks. In addition, the survey

found that poverty and occupational status were strongly associated with the levels of literacy attained (Barton 1994). SES is not the whole story, however. Camara and Schmidt (1999:8) state that "[M]iddle SES white children are more likely to be at the highest proficiency levels of NALS than upper-SES Hispanic and African-American students. Such patterns are also found on nontest measures such as grades and class rank". They point to the need for further research into the causes of test score differences between groups.

2.8.2 Medium of Instruction

It is by no means an unusual situation to learn through the medium of a language other than one's first language. Children in India, most countries in Africa, China, the United States, Canada, and many other countries attend schools where the language used is not their first language, or is a significantly different variety. In South Africa, the official government policy on languages is one that puts a premium on multi-lingualism, and on the role that learners' home languages can and should play in promoting and supporting learning. The additive bilingual model that forms the basis for government thinking on language in education begins with the learners' home language, and then (as its name implies), adds other languages in subsequent years. In this model the use of one language (the first language) is not abruptly discontinued as is the common practice in South African schools, but is maintained, although only as one of two or more languages which are used as 'languages of learning' (DoE 1997c). At present, however, parental demand and public perception appear to support the use of English, either to replace the home language at around the fifth year of schooling, or to introduce it in the first year as the medium of instruction. This issue, and its implications for how learners know and understand, is taken further in Chapters Five and Six.

2.8.3 School Quality

Findings on the impact of school quality are somewhat difficult to interpret, with some researchers claiming surprisingly little effect (e.g. Coleman et al 1966, cited in Muller & Roberts 2000). However, what is generally accepted is that in poor communities, with illiterate or poorly educated adults, the quality of the school has a greater influence on the life chances of a learner than it has

in more privileged communities (Crouch & Mabogoane 1998, Muller & Roberts 2000). In such a context, if schools are of poor quality, learners are further disadvantaged.

2.8.4 Stereotype Threat

In the United States, data from the College Board (Camara & Schmidt 1999, The College Board 1997) suggest that regardless of economic or educational background, students who are black, Latino, or American Indian, do not perform as well as their white or Asian counterparts on tests such as the SAT. Several explanations are put forward for this gap: the low expectations that teachers at all levels have of minority students; educational programmes that focus on minimum competency levels and thus implicitly set these minima up as standards; the strong perception in minority communities that academic success is "acting white"; and learning materials that do not reflect the experiences of minority students (Weissert 1999:A42).

The effects of low expectations are stressed by Garcia, Jorgensen and Ormsby (1999), who recount the outcomes of controlled studies of test taking motivation in Stanford University students, conducted by Steele and Aronson in 1995. In these studies, African-American and white students wrote two tests, one of which they believed to be 'reflective of' the Graduate Record Examination (a high-stakes admissions test used by many prestigious graduate schools in the United States), and the other to be of no account. In the latter test, the two groups performed at a similar level – on the former, however, African-American students scored significantly lower. On the basis of evidence from this study, the researchers conclude that "... students could very well internalise a stereotyped image of themselves when put into judgement situations" (Garcia et al 1999:7).

In South Africa, where educationally disadvantaged students are in the majority, the phenomenon of stereotype threat nevertheless still applies. Transformation in all spheres of public life is rapidly being implemented: in educational life, however, particularly at selective tertiary institutions and at fee-paying schools, there are still few black role models. The great majority of black students achieve results in the lower ranges (50 – 55%), and there are very few black (or women) professors.

2.8.5 Sex and Geographical Origin

Discrimination on the basis of sex and geographical origin are two other sources of educational disadvantage, the results of which can be clearly seen in the preponderance of women in the discipline areas of education and humanities, and the very poor Senior Certificate results of rural ex-DET schools which limit access opportunities for those rural students.

While it is possible for individuals to overcome the disadvantaging effects of one or even two or three of the factors listed above, it is virtually impossible to overcome all of them. Learners whose parents are poor and illiterate but who have the good fortune to attend a school of reasonable quality (such as a mission or church school, or a school paid for by a philanthropic organisation or a parent's employer) will not experience the same degree of disadvantage as that of a student from a similar family background who attends a poorly run, largely dysfunctional school. However, in the South African situation, the legacy of apartheid is such that the overwhelming majority of black students come from low SES homes, and attend schools of poor quality. In addition, the medium of instruction is not their first language, and there are few successful professional role models in the communities in which they live. Upward mobility amongst blacks, as for any other group, is marked by a move into safer, quieter neighbourhoods, with the consequential further educational impoverishment of the communities in which the majority still live.

Educational disadvantage in South Africa is still largely arranged on racial lines. However, the rapid expansion of the black middle class, and the dramatically increasing number of black students attending previously whites-only schools (which charge substantial fees and provide markedly better learning opportunities than those offered by former DET schools), has eroded the sharp delineation of educational privilege along colour - although not class - lines. This can be illustrated by the UCT situation, where the numbers of black students at the institution are increasing while the numbers of ex-DET applicants and registered students are decreasing (University of Cape Town 2001). For a selective institution such as UCT, it is thus not correct to assume that all black students are educationally disadvantaged, and so some indicator other than 'race' is necessary.

The level of fine-tuning of such an indicator or indicators is highly context-dependent – it is, as Zaaiman (1998:23) suggests, "...a relative concept". At the University of the North, for example, the locale for Zaaiman's study, the student body was 100% African (Bot 1998), and, while precise numbers are hard to come by, it is safe to assume that the overwhelming majority of these students would have attended schools previously administered by the DET. In this case, the use of "ex-DET" as an indicator is not helpful, and it is necessary to develop finer grained sieves.

UCT, which Klitgaard would term a "right-tail" institution in that selection occurs largely from the right tail of the Senior Certificate results range, has an extremely heterogeneous applicant pool. Indeed, in 2000, 47% of its first-year student body had A or B SC aggregates, and, over the period 1996 – 2000, 84% of its 1999 applicants came from other than ex-DET schools (UCT 2001). In this case, unlike the situation at the University of the North, the ex-DET indicator serves to identify the most educationally disadvantaged group. This was illustrated in Table 2.2 above.

Chapter Six (section 6.3.1) returns to the issue of educational disadvantage, although from the perspective of the impact of educational disadvantage on knowing and learning, and the implications of this for assessment.

2.9 Effectiveness and Efficiency

Coombs (1994:286) suggests that "... efficacy encourages us to distribute [educational] resources in such a way as to maximise the total talent available to society ...". One interpretation of this is that those who have "... greater natural capacities for achieving valuable knowledge, abilities and skills ..." (op cit:286) ought to be given greater access to and hence share of these resources, as they are more likely to be able to contribute constructively to society. However, distribution on this basis is likely to perpetuate social inequities, in that those with developed or developing skills and talents are likely to be those who have had advantageous educational opportunities, or access to educational resources. Green (1971, cited in Coombs 1994) calls this the 'durable injustice'.

Another view on how to maximise the talent available to society is to identify those groups most oppressed by society, and to target scarce resources at them, thereby bringing them into the talent

pool. This long-term view has clear merits, as it stands to reason that much potential talent is wasted in societies with large numbers of poorly educated, poor people. However, using Table 2.2 above as an illustration, it would mean that ex-HoA applicants with scores of 68% (for example) would not be offered a place on the basis of their test performance even though it is highly likely that they would be academically successful at the institution, and would require fewer educational resources (in the way of foundational or extended programmes), whereas ex-DET applicants with a lower score and who would require significant educational resources would be offered places. The identification of ex-DET students as deserving of a place is on the grounds of relative merit (they were in the top 20% of their candidate pool) and not on developed ability. Maximising talent through resource distribution is thus a long-term, expensive option, but is one of the few viable equitable options in societies where access to educational resources has been skewed.

If one accepts the need for a diminishment of efficiency on the grounds of equity and long-term good, it is likely that one has to accept a reduction in short-term selection efficiency as well, as there is no doubt that simply taking school-leaving scores as the sole basis for admissions is the easiest and cheapest method. The Higher Education White Paper (DoE 1997b:1.22) states, with regard to effectiveness and efficiency, that:

> "An effective system or institution functions in such a way that it leads to desired outcomes or achieves desired objectives. An efficient system or institution is one that works well, without unnecessary duplication or waste, and within the bounds of affordability and sustainability."

In this view, much depends on the interpretation of the word 'unnecessary'. If it can be shown that it is necessary, in terms of equity and the goals of having an excellent and diverse student body, to develop and/or use more than one set of indicators of merit, then this constraint falls away. The 'bounds of affordability and sustainability', too, need to be interrogated: it is cheaper, granted, to simply use Senior Certificate results (or to rely on bureaucratic bungling), if what one is costing is the process of selection. The costs to the state (estimated to be in the region of R1,3 billion per annum (DoE 2001)), the institution, the student, and the educational process, of high failure rates and the impossibility of a planned approach are, however, far higher than those associated with more expensive initial selection costs.

Clearly, what is needed is a compromise.

According to Zaaiman (1998:31), "The main aim of selection to higher education is to identify students who will succeed in a specific academic programme". This identification, of course, represents an ideal situation, but is hardly practical when each institution has numerous programmes. This is not to say that a fine sieve is not necessary at some stage of the selection process – in any case, the information yielded by an entry assessment is invaluable for good course design – but that there is a need for 'broad spectrum' approaches as well as programme-specific procedures.

2.10 Conclusion

This chapter has argued that selection for Higher Education is not only inevitable, in that the alternatives are either unrealistic or unfeasible or both, but that the responsible use of taxpayers' money to meet society's needs demands that the most capable applicants should be admitted. Developing fair and feasible selection procedures, however, is a complex matter in a society as divided and inequitable as that of South Africa, where the effects of the apartheid education system are still powerfully experienced. Undue reliance on school leaving examination results will unfairly discriminate against candidates whose schooling has been of poor quality. At the same time, abandoning the Senior Certificate, or seriously downgrading its use, will cause incalculable harm to the school system, and will impoverish the information about candidates' curriculum-aligned knowledge and skills. Later chapters in this study investigate various assessment options which could be used in combination with SC results, based on the assumption that this combined use would be both more equitable in terms of not simply advantaging further the already advantaged, and would also provide additional, qualitatively different information about candidates, thereby enriching the selection process.

CHAPTER THREE
GENERAL ASSESSMENT ISSUES

3.1 Introduction

Chapter Two outlined the major options open to institutions in constituting their student bodies, and discussed many of the consequential implications of these choices. The discussion concluded that some form of selection was essential (even if only, in some contexts, for high demand curricula), and that one of the most feasible bases for selection was academic merit. Academic merit, however, is highly influenced by the prior educational opportunities experienced by individuals, and so equitable selection procedures must go beyond the testing of achievement. It was argued that achievement tests, such as the Senior Certificate examination system in South Africa, yield important information about what candidates bring with them to the next stage of learning, as they are closely tied to what students have been taught. For students whose learning opportunities have been sub-optimal, however, the SC examination ought not to be the only source of

information about academic merit, and Chapter Two concluded that other assessment options should be investigated with the aim of establishing such additional sources of information.

One of the difficulties faced by those working in the area of assessment is the widespread anxiety and distrust the topic arouses. This distrust is certainly not unique to South Africa, but given the very recent history of systematic educational discrimination, it is particularly acutely manifested here. This is not surprising, since in nearly all high-stakes examinations - such as the Senior Certificate - the results of black South Africans tend to reflect unequal educational opportunities, and thus to be considerably lower than those of their still relatively advantaged white peers.

Much of the suspicion and hostility attracted by assessment, however, is due to confusion about the aims, and limitations, of various types of assessment. Chapter Three includes a discussion of major issues surrounding assessment in general, including a synopsis of the origins of large-scale testing systems, a brief treatment of various types of tests, and an overview of issues relating to test fairness. The choice of issues is selective, made on the basis of their relevance to the study rather than of comprehensive coverage of the field.

The term 'assessment' covers a wide range of meanings. Webster's New Collegiate Dictionary (1977), for example, defines "to assess" as follows:

> "1: to determine the rate or amount of (*as a tax*) 2: to impose (*as a tax*) according to an established rate 3: to make an official valuation of ... 4: to determine the importance, size, or value of ..."

The last meaning, which includes the making of a judgement, illustrates the large range of meanings which can be attached to, or derived from, use of the term 'assessment'. One can assess a painting as being beautiful, authentic, or valuable - all very different kinds of undertaking. Narrowing the term to 'educational assessment', however, still includes a wide array of meanings. At one extreme, the term could refer to an informal, five-minute quiz at the end of a classroom lesson, designed by the teacher as a spot check on the effectiveness of the lesson. At the other extreme, it could refer to the tests being produced by the 'TOEFL 2000' team at the Educational

Testing Service (ETS), involving numerous research teams, test trials, hundreds of thousands of test takers, and important long-term consequences for the test takers.

3.2 Types of Tests

3.2.1 Intelligence Tests

Traditional approaches to intelligence testing are based on the concept of mental age, usually attributed to Alfred Binet, who developed the idea of measuring intelligence in the early 1900s. An individual's intelligence quotient (IQ) is computed by using a standardised intelligence test to measure mental age, multiplying this age by 100 and dividing by the individual's chronological age.

Theories of intelligence vary widely, although almost all are concerned to a lesser or greater extent with individuals' ability to adapt to new situations, to learn, to deal with abstract symbols and relationships, and to solve widely differing, and new problems. In very general terms, intelligence may be thought of as "... a hypothetical property that people possess to some degree according to their ability to solve diverse problems of differing degrees of complexity and novelty" (van den Berg 1989:88). Immediate problems arise even with this definition, however, as degrees of complexity, and notions of what counts as novel, are controversial. Tests of intelligence are thus vulnerable to accusations of unfairness and bias, unless the population being tested has a very high degree of homogeneity in terms of biographical factors, such as age, education, socio-economic status, ethnicity, and language. In the absence of convincing evidence on the usefulness and validity of tests of intelligence in the South African context, with its very low degree of heterogeneity in terms of these factors, such tests are not considered appropriate for use in selection procedures for Higher Education.

3.2.2 Achievement Tests

Achievement tests aim to assess mastery of a specified body of knowledge. The Biology examination in the Senior Certificate set is an achievement test, as it is based on the Biology syllabus. A complicating factor is of course whether, if some candidates have not had the opportunity to cover the material in the syllabus, the Biology paper is in fact an achievement test

for those candidates. This issue, crucial in an inequitable context such as South Africa, is addressed in Section 3.4 below.

3.2.3 Aptitude Tests

Aptitude is one of the most poorly defined concepts in psychological theory. As Snow (1992:5) suggests, "Aptitude is an old term for an old concept still widely used, but also widely misused and misunderstood in much scientific, professional, and public parlance today". Much of the confusion can be attributed to different views on the influence of heredity and environment, with some theorists suggesting that an aptitude is innate while others highlight the role of environmental interactions (e.g. Snow 1992, Anastasi 1982, and Rogoff 1984).

From the perspective of testing, the former view, viz. that aptitude is innate, would tend to favour testing approaches that downplay the impact of prior educational or cultural experiences. The latter view, on the other hand, would tend to favour a more dynamic approach which views achievement as "... dynamically determined by educational and environmental interaction" (Browne-Miller 1995:169). Chapter Four (section 4.4) explores the implications of these views in some detail.

The 'common sense' view of aptitude is that it refers to learning potential. Thus, one might say of an individual that 'she has a real aptitude for learning the violin', meaning that she seems to find it easy to learn to play the violin, as opposed to another pupil who struggles with the violin but appears to sail through any problems involving numbers. One point on which there is general agreement is that aptitude always implies prediction by means of some increased likeliness of prospective outcomes. When one talks of aptitude one does so by specifying a particular setting, as in "an aptitude for x, or for y". Thus, as Snow (1992) suggests, part of defining an aptitude is defining the situation.

Tests such as the SAT1: Verbal Reasoning Test[16] are based on the concept of scholastic aptitude. This concept, according to Donlon (1984:125), refers to "... general abilities that are developed over a number of years, influenced by experience both in and out of school ...". It is the relatively static nature of these abilities in this understanding of aptitude that is the distinguishing feature. Aptitude tests of this kind tap general skills which, while at least partially dependent on prior learning opportunities, are somewhat abstracted from any particular curriculum. They are similar in many respects to intelligence tests, although they reflect the learning and performance demands of a particular situation (e.g. college level work). In other words, they define the situation for which they are considered relevant, as was suggested above. For this reason, they are often perceived as more defensible and legitimate than are intelligence tests.

3.2.4 Proficiency and Placement Tests

Proficiency tests share common features with both achievement and aptitude tests. They differ in important respects from both, however. In relation to the former, they differ in that they do not refer to a specifically defined course of instruction. In relation to the latter, they differ in that they are designed to yield information about present levels of skills and knowledge in terms of a future situation and purpose: aptitude tests aim rather to assess the ability of an individual to acquire new skills and knowledge. `Proficient' in this sense means having sufficient command of some skill or knowledge for a particular purpose, and as such tests of this kind usually require some kind of needs analysis to be undertaken during the design stage. An example of this analysis could be a specified level of English language proficiency for successful study at an English-medium institution. For a university admissions test, this purpose must include the successful completion of representative academic tasks, such as reading, listening, laboratory work, note/essay/report writing, and so forth. In this sense the criterion situation (performance at college) defines what counts as proficient and thus what needs to be tested.

[16] The SAT was originally called the Scholastic Aptitude Test, then for a brief time the Student Assessment Test, but is now known officially by its initials (Hoff 1999). These shifts reflect public confusion and suspicion about such testing.

In terms of language testing, proficiency tests are usually tests of foreign or second language proficiency, where the level of language mastery (e.g. beginners, intermediate, advanced) is assessed. Large-scale examples of such tests used internationally are the TOEFL and the IELTS.

Placement tests are used, as their name implies, to place students into appropriate courses or curricula. Because their curricular links are so important, placement tests are usually institution-specific, or, if they are standardised tests developed elsewhere, can have institution-specific cut-off points for placement purposes. Placement decisions can be based on diagnostic tests, used to identify strengths and weaknesses in particular areas, achievement tests, or proficiency tests.

3.2.5 Curriculum-Aligned versus Non-Curriculum-Aligned Tests

Curriculum-aligned tests are tests which are based on a preceding course of instruction, whereas non-curriculum-aligned tests are generally based on a notion of core skills and knowledge in a specific domain. The two kinds of tests have different strengths and uses, as discussed below, but both support the argument that high-stakes examinations can have a crucial role in motivating students – not only to complete schooling, but to exert some effort. This role is one of the primary motivations for retaining and strengthening the school-leaving examination in South Africa.

Heyneman and Ransom (1992) set out some of the tensions in this area, which they characterise as 'coverage versus fairness'. Clearly, if a test is meant to provide information on the quality of teaching and learning in schools, and the extent to which students have mastered the syllabuses in different subject areas, it is essential that the tests be closely aligned to the curriculum. However, in a context where wide disparity exists in the quality of educational provision and thus in the extent to which the curriculum is likely to have been covered, tests may not reflect what has been learned as much as who had an opportunity to learn. On the other hand, however, setting tests only on those parts of the curriculum likely to have been covered by the majority of learners will provide a truer reflection of what has been learned, but will not accurately measure what the curriculum was intended to achieve.

Heyneman and Ransom conclude that the appropriate response in such a situation is not to move away from curriculum-aligned exams, but to use examinations to ensure that greater coverage of the curriculum is achieved. In other words, they endorse a measurement-driven instruction approach, and argue that balanced examinations will encourage and promote teaching and learning. However, at the same time they caution about the unfairness of assuming that all candidates would have the same opportunities in terms of preparation.

This view points to the need to have two kinds of tests:

- Achievement tests which are based on the curriculum and will allow conclusions to be drawn about the extent to which effective teaching and learning have taken place, and
- External, standardised tests which are not tied to the curriculum of the schools, and which therefore provide students from poorer quality schools with an opportunity to demonstrate talent.

These two kinds of tests, used in combination, form the basis for much of the argument in this study.

3.3 Origins and Functions of External Standardised Testing

Standardised tests are tests "... designed to provide all test takers with a uniform experience and an equal opportunity to demonstrate the skills or knowledge being measured" (Beatty, Greenwood & Linn 1999:3). More specifically, they have the following characteristics:

- They are expertly constructed, usually with trial testing, analysis and revision. Yardley (2000:32) wryly observes: "How is a standardised test made? In a politically charged environment, the quick answer is: Very carefully". As trial testing is not always possible, however, previous versions of tests are sometimes used as the basis for test development. Reliance on previous versions is particularly the case with achievement tests, such as the Senior Certificate examination, where it is clearly not possible, for a variety of reasons (not least of which is security), to conduct adequate trials. In such cases, the trial testing, analysis

and revision is sometimes conducted on a *post-hoc* basis, or achieved through statistical manipulations such as 'norming'.

- They include explicit instructions for standard administration and scoring.
- They may provide tables of norms for score interpretation purposes, derived from prior administration of the test to a sample of candidates.
- They can be used for making score comparisons amongst candidates who have taken the test under the prescribed conditions.

However, the fact that a test has been standardised does not mean that all candidates have "... an equal opportunity to demonstrate the skills or knowledge being measured" as implied above in Beatty et al's definition of standardised tests. It does, however, mean that all due care will be taken to eliminate bias. This difficulty, and the controversies to which it gives rise, are discussed in section 3.4 below.

A common misunderstanding concerning standardised tests is that they refer only to tests utilising particular formats (for example, multiple-choice format tests). However, almost all tests could be standardised if the effort was thought justifiable i.e. if it was necessary to compare scores across tests. The tendency for standardised tests to rely on multiple-choice items is because they cut out a large source of potential variance in the marking procedures, and thus are easier to standardise.

Because of the expertise required in the construction of standardised tests, it is usual that they are also external tests – that is, tests developed and administered by an agency (or individual) which is not connected to the instructional process. However, such tests need not be external. For example, a school might have developed a scholarship examination that has, over the years of its use, been standardised – and might routinely admit and support all candidates who achieve over a certain cut-off point. In this case the test is standardised but not external. Conversely, an institution might employ an external agency (or individual) to develop tests for its use, which might or might not be standardised: or the institution might use a standardised, external test in ways which invalidate the standardisation requirements. On the whole, though, external tests tend to be standardised, and vice versa.

In understanding the power exerted by external, standardised testing initiatives, it is useful to consider some of the reasons for the origins of external examinations, and to reflect on the extent to which these factors still hold. Eckstein and Noah (1993) suggest the following broad categories, adapted here for use as a framework for discussion.

3.3.1 Replacement of Monopolies of Birth and Wealth, and Containment of Corruption

As far back as the Han dynasty in China (201 BC to 8 AD), external, standardised examinations were introduced in order to replace patronage as a means of allocating positions in the civil service and of selecting its mandarin class (Spolsky 1995, Eckstein & Noah 1993, 1992). In other countries, however, examinations have only relatively recently been used as a technology for social engineering. Only in the mid to late nineteenth century, for example, was the practice of buying and selling positions in the British government service abolished, and replaced by examinations. This change represented an advance on the previous practice, but did little to offset the advantages conveyed by birth and privilege. In this respect, Amano (1990) describes how, after competitive examinations were introduced in Japan, the previously advantaged Samurai class benefited greatly from their introduction, occupying disproportionately large numbers of places at prestigious higher education institutions.

Similarly, the SAT 1: Verbal Reasoning Test, developed in the early twentieth century in the United States, had its origins in a desire to create a classless, democratic society. Ironically, however, "... a device meant to eliminate an American class system has instead helped create a new one" (Lemann 1999a:53). In *The Big Test: The Secret History of the American Meritocracy*, Lemann (1999b) traces the growth of the SAT, and concludes that "... the project of picking just the right elite and the project of building the perfect domestic society turn out not to be very closely related" (1999a:56).

The SAT is widely criticised, for example, for the persistent and substantial black-white gap in average test scores, as well as a less dramatic but still significant gap between the average test scores of male and female candidates. However, as is pointed out by The College Board (1999),

the SAT gaps in group performance mirror those identified in reading, writing and maths standardised tests administered at other stages in the educational process (e.g. in the fourth, eighth and twelfth grades, on admission to undergraduate colleges, and at the postgraduate level).

Nevertheless, decisions made solely on the basis of SAT scores tend to reinforce and perpetuate the *status quo*[17]. Tests based on the assumption – evident in the arguments of the early developers of the SAT tests - that merit and ability are somehow uninfluenced by, and/or independent of prior life and educational experiences - will inevitably have this consequence (Lemann 1999a&b, Crouse & Trusheim 1988). This position of 'qualified individualism' (Zaaiman 1998) and its implications for fairness in selection was discussed in Chapter Two, section 2.7.2.

Claims that a test such as the SAT can on its own radically alter the pecking order in a society are clearly unfounded, as demonstrated by the American experience. Nevertheless, the introduction of examinations has had, and continues to have, an effect on the ways in which societies structure themselves. As Foster (1992:123) suggests, "Examinations may sometimes exert a deleterious effect on pedagogical practice and learning outcomes, but in terms of their role as instruments of status allocation they may be infinitely preferable to the other alternatives".

3.3.2 Promotion and Support of Learning and Teaching

One explanation for the origins of external examinations is the belief in their use as an instrument to improve teaching and learning processes. The relationships between teaching and learning on the one hand, and the assessment of teaching and learning on the other, are extremely complex even when assessment is carried out as an integral part of the teaching/learning process. When assessment is conducted externally in an attempt to influence these processes, the relationships become even more complex. The discussion below focuses on some of the arguments and evidence relating to the role of external assessment procedures in learning and teaching. The discussion is arranged in terms of the role of external examinations in performing the following

[17] An example of how this could be achieved is the ETS Strivers Study, which set out to identify (through a complex regression analysis procedure) candidates who scored at least 200 points higher than their estimated SAT score. The estimations were based on fourteen different categories, including family background and income and parents' educational level (ETS 1999). However, the public outcry at the notion that the actual score obtained by an individual

functions: raising levels of knowledge and skill (3.3.2.1); and monitoring and improving the effectiveness of teachers and schools (3.3.2.2).

3.3.2.1 Raising Levels of Knowledge and Skill

A basic and persistent belief of the assessment-led-instruction school of thought is that good assessments will promote good teaching and learning, and thus help to raise levels of knowledge and skills. In a discussion on the role of motivation in school performance, Kellaghan, Madaus and Raczek (1996) summarise the main arguments used by advocates of the 'measurement-driven instruction' school of thought. According to the arguments, students (and teachers) will work hard, and thus general standards will rise, if:

- clear and demanding goals and/or standards are set, and operationalised in examinations. Essentially, the assumption is that examinations should clearly reflect the goals to which students have been working – so that if, for example, 'problem-solving' has been a goal, the examination should contain items which unambiguously tap this ability or skill. As is clear from the extensive literature (e.g. Jones 1997, Shepard 1997, Green 1995) on this topic, however, it is extremely difficult to develop items which convincingly and legitimately tap a particular skill. Nevertheless, the point is well made that examinations which bear little discernible relationship to the goals of an instructional unit will certainly have a demotivating effect on learners.
- attainment levels of students in terms of these standards are clearly visible to students via their performance on examinations (i.e. their performance is clearly interpretable in terms of the goals/standards). Relating performance on examinations to the way that results are reported is also undeniably linked to motivation: what can a candidate understand about what she knows or can do in a particular learning area when she is given an "E"? Norm referenced tests yield relative information, not information in terms of learning goals or standards, unless great care is taken to provide meaningful descriptors for bands of scores, for example.
- connections between effort at school (and schoolwork generally), exam performance, and future life chances are clear. The negative effects on learners of a view of assessment as no

might be tampered with on what could appear to be affirmative action grounds forced the ETS developers to retreat, and it is at this stage not clear what the future of the Strivers project is.

more nor less than a lottery are well documented, and have undoubtedly contributed to a lack of motivation to make efforts in what are often adverse learning conditions. As the DoE (1998) Report states,

> "... a disturbingly high proportion of interested parties views the Senior Certificate as a lottery: how you do is arbitrary and unrelated to either effort or circumstances. It has been suggested, too, that this view of the SC is subtly reinforced by some teachers who are not keen to see their learners' poor performance ascribed to inadequate syllabus coverage or teaching, and find it convenient to lay the blame on 'the examination" (op cit:47).

- incentives and disincentives are tied to exam performance.
- assessments are compulsory and take place at clearly defined stages in schooling.

It needs to be recognised, however, that the role of motivation is more complex than is implied above. Fundamentally, for an individual to be motivated, s/he should 'buy into' a goal. However, in schooling as in other arenas, it is by no means certain that what lures one learner will lure another. Nevertheless, ensuring that the curriculum and assessment are closely aligned appears to be essential if the distorting effects of external assessment are to be minimised.

In strong contrast to the assessment-led instruction approach, many educationists and others argue that measurement-driven instruction has a pernicious effect on the curriculum. Madaus, for example, views this approach as "... nothing more than psychometric imperialism" (1988:84). This criticism should not be taken to read that he rejects the value of assessment *per se*, but rather that he believes it should serve the curriculum, not drive it. In an article written at a time in the United States when proposals to develop common examinations for each of the secondary school curricular areas were being formulated, Madaus (op cit) derived seven principles, based on the findings of studies on the power of testing to influence the curriculum. Three of these which are particularly relevant to this study are discussed below.

The first principle asserts that the power of tests derives entirely from how important they are perceived to be by test users and takers – it is, he argues, a 'perceptual phenomenon'. This principle includes the power of tests to mean or stand for various phenomena. In this way, Madaus states, "[T]est results become a synecdoche for standards" (1988:89). Thus, when test scores

rise, it is believed that standards in that area have risen. However, as is illustrated below, this change could as easily mean that candidates have become more adept at taking the test, or that the results have been manipulated in some way, as that they are more proficient in the area being assessed.

In South Africa, this presumption can be seen very clearly in the publicity surrounding pass-rates in the Senior Certificate examination. If they rise, standards are believed to have risen, and vice versa. The controversy surrounding the 2000 SC results are a case in point, with government officials claiming improvements in schooling (Grey 2001:2), and others such as noted local educationist Jonathan Jansen stating that the nine-percentage point increase reported for the 2000 results was not credible (Monare 2001).

A further principle put forward by Madaus states that if a test has important perceived consequences, it will be taught to. In and of itself, of course, this need not be a bad thing. If the test is measuring skills which are agreed, through some legitimate process, to be important, then instruction aimed at teaching these skills ought to be beneficial. This argument is at the heart of the proponents of assessment-led reform. The challenge, however - which such proponents would acknowledge - is to develop and maintain methods of assessment that authentically represent the skills. The extent to which the skills and the measurements do not correspond determines the extent to which teaching to the test will be negative in its consequences.

An illustration of how scores can change without there being a corresponding, generalized change in knowledge is provided by Shepard (1997). As she states, students "... can appear to know when they don't really know" (op cit:8). The example is based on the performances of 9 year-olds in a high-stakes testing district, i.e. a testing district where accountability testing was exerting pressure on schools to raise scores, and is derived from a study by Koretz (1991).

STANDARDISED TEST ITEM

Instruction	Options
87 -24	A 63 B 53 C 64 D 62

ALTERNATIVE TEST ITEM

Instruction	Options
Subtract 24 from 87	A 63 B 53 C 64 D 62

As can be seen, the vertical subtraction problem of the standardised test item was changed to a horizontal problem in the alternative test version. In both cases, however, candidates were required to manipulate the same two numbers, 87 and 24, and in both cases the same multiple-choice options were supplied. However, the performance of the candidates changed depending on which format was presented. 83% of the candidates got the standardised test item right, whereas only 66% of the candidates got the alternative test version right.

This difference in performance raises serious questions about the 'robustness' of the mastery of the skill. As Shepard suggests, students can appear to understand, "... if we keep asking them to demonstrate skills in exactly the same format" (1997:20). As this example demonstrates, however, changing the format can change performance levels. It also raises an important issue related to the existence of 'generic' skills, as illustrated below.

The issues are complex. If a candidate can perform similar tasks in one context but not in another - or in one format but not another - it is not clear what this performance means. It could, for example, mean that the candidate's understanding or knowledge is incomplete - that s/he has learned the format strategy but does not have a sound understanding of the skill. On the other hand, it could mean that one of the contexts contains inadequate or misleading items/item formats which prevent her/him from being able to demonstrate mastery.

The difficulty of explaining why there is a difference in scores when different formats are used, however, does not lessen the importance of the existence of this difference. What this evidence points to is the need for the developers of assessment procedures and instruments to insist on the inclusion of as wide a variety of formats as possible. Equally important, however, is that the dependence of performance on format and context is not just a measurement problem. It is, rather, the responsibility of all involved in the educational process to ensure that every avenue possible is used to "... ensure transfer and generalized knowledge" (Shepard 1997:27), and this imperative includes the avenue of assessment. A candidate ought not to be assessed as proficient until s/he has demonstrated mastery in more than one context.

The third principle discussed here asserts that irrespective of context, wherever a high-stakes test is established, "... a tradition of past exams develops, which eventually de facto defines the curriculum." (Madaus 1988:93). The relationship of the syllabus to the examination is crucial in this regard. If there is conflict between the aims of the syllabus and the way the course is assessed, the examination's implicit curriculum will win. One of the difficulties is of course that tests have to sample what is in a syllabus, as complete coverage is not possible. However, because some parts of the syllabus are more 'testable' than are others, these parts will tend to appear more frequently, and thus will assume a more important place in the curriculum than was originally intended. Understandably, in a high-stakes situation, teachers try to maximise overlap between what they teach and what is examined - a strategy referred to by Resnick and Resnick (1992:57) as 'curriculum alignment'. Generally, scores rise as the curriculum is aligned to the tests, i.e. where there is increasing overlap. In situations where syllabuses are not available, or where confusion exists about which version of a syllabus is in force[18], of course, it is easy to see how examinations begin to define the curriculum.

A further principle that might be advanced is the tendency for external examinations to encourage an "us and them" climate in the classroom. By characterising the examination as the enemy, teachers and students can be drawn together, sharing helplessness/hostility. As Eckstein and Noah (1993:220) suggest, "students and teachers can become allies in the business of outsmarting examiners". In addition, teachers (and learners) who are not conscientious can use this distrust as a way of explaining and excusing poor performance. Such collusive practices may be one of the reasons why learners and parents allow negligent teachers to escape public blame for disastrous examination results.

From the discussion above, it can be seen that using external examinations to raise the level of knowledge and skills in an educational system is an extremely complex undertaking. In order to minimise negative consequences, great care must be taken, before such an undertaking is embarked upon, to ensure systematic and in-depth consultation with teachers in the design,

[18] Such confusion was found in the investigation of the Ministerial Committee looking into the Senior Certificate examination in 1998, where teachers and regional coordinators were using different syllabuses for ESL-HG, and where not even the National Department could produce copies of the most recent form of the syllabus (DoE 1998).

content, and administration of the tests. In turn, however, such consultation requires a well-trained and highly motivated teacher corps, as is the case with the Queensland Core Skills Test scheme in Australia[19]. In addition, a variety of approaches and formats is essential if the assessment initiative is not to distort the aims it was established to achieve. Lastly, if proficiency tests are to be introduced, as it is argued in later chapters in this study is essential in contexts of great educational disparities in provision, their use should be in addition to, and not as an alternative to, that of achievement tests.

3.3.2.2 Monitoring and Improving the Effectiveness of Teachers and Schools

A further function that external examinations were intended to perform in terms of promoting and supporting learning and teaching was that of providing a means by which progress and impact could be monitored. In the United States, the establishment of the National Assessment of Educational Progress (NAEP) project in the early 1960s signalled the beginning of the systematic use of test data in that country to describe the status of education at State level. In South Africa, the Human Sciences Research Council (HSRC) was responsible for much large-scale survey research, often based on the use of standardised tests.

Some of the reasons for this move to collect test-based data on performance include:

- A concern with outcomes – what are learners actually emerging with at the end of lengthy and expensive courses of instruction? And how are these outcomes related to inputs of various kinds?
- A concern with identifying problem areas of under-performance so that compensatory programmes could be put in place.
- A need to monitor the effectiveness of various curriculum development and compensatory programmes.

The use of external tests as administrative mechanisms in policy relates to their use as mechanisms of power: strong or weak performance on the tests would have good or punitive

[19] This assessment regime is discussed in Chapter Five.

consequences. For example, regions within a province that perform very poorly could, as in the States, be required to submit and implement once approved, a remediation plan. There is widespread evidence in South Africa of such use of Senior Certificate results, with schools that have achieved 0% pass rates on the examination being targeted for remedial action. The increased 2000 SC examination pass rates are being attributed at least partly to such initiatives (Grey 2001:2).

It is clear that the existence of test data on performance at exit levels is of great benefit to policymakers, and can act as a safeguard for learners who are trapped in dysfunctional schools or learning situations. The data can provide evidence of problem areas and thus can be used by educational reformers as a basis for the establishment of reform measures. As has been discussed above, however, there are several serious problems associated with the use of tests as providers of information. These problems are likely to be exacerbated in 'high-stakes' conditions, unless – as was suggested above - steps are taken to minimise them.

3.3.3 Selection and certification

The rise in influence of external examinations is closely connected to their use in selection and certification procedures. As Beatty et al (1999) suggest, the sorting process that takes place at various points in education requires "... an efficient source of comparative information" (op cit:1). At this stage, there appear to be no substitutes for tests, which can deliver relatively low-cost information. Given the range of quality across educational systems, standardised tests provide a means by which institutions can discriminate amongst thousands of applications. Chapter Four of this study lays out some of the requirements for testing arising from the need to select.

3.4 Appropriate Test Use

The three principal criteria adopted by the United States Committee on Appropriate Test Use (Heubert & Hauser 1999) to assess responsible test use are:

- Measurement validity (whether a test is valid for a particular purpose, and whether it accurately assesses a candidate's knowledge);
- Attribution of cause - whether a candidate's test performance is due to poor educational preparation, some feature of the test itself (such as confusing instructions, language issues), or irrelevant test content; and
- Effectiveness of treatment (some have called this consequential validity, as it focuses on what follows the test, and the validity of this consequence).

These three criteria are highly interrelated, in that problems in one area will inevitably compromise the other. Nevertheless, their separation is useful for the purposes of discussion.

3.4.1 Measurement Validity

Measurement validity is central to the activity of testing, as it addresses the question of the meanings that can be drawn from the results. Validity refers to the extent to which one can be confident that what a test sets out to assess is in fact assessed. Various kinds of validity are distinguished, such as construct, content, face, concurrent, predictive, and consequential validities. These kinds of validity are discussed in some detail in Chapters Eight to Twelve, and are thus not elaborated on here.

Resnick and Resnick (1992) suggest that the contrasts between the different functions of testing "... argue for a public discourse on assessment that maintains a clear distinction ..." (op cit: 51) between them. Essentially, they insist that all of the different functions have different implications for and relationships with teaching and curricula, require different assessment procedures and systems, and are aimed at different audiences.

For example, tests that are designed to yield information about individual candidates (e.g. the SAT tests in the United States, the SC in South Africa, the IELTS in the United Kingdom) are only valid if the tests can be said to be fair for all the candidates. In this case, the audience includes the candidates, as well as future employers or Higher Education admissions officers. Tests aimed at addressing the monitoring and accountability needs of a system, however, might have as a

fundamental target the identification of discriminatory educational practices and poorly functioning schools or areas (e.g. NAEP in the United States) and so would be regarded as valid even if the tests were not 'fair' for all candidates. Here the audience is the authorities charged with managing the educational system, at whatever level (state, district, school), as well as the public at large.

In the former case, the tests would need to be closely aligned to preceding curricula, as is the case with achievement tests such as the SC in South Africa. In the latter, they would need to be carefully and uniformly non-aligned, as can be seen in tests such as the SAT. The difficulty arises when tests of one kind are used as though they were of the other. Examples of the confusion that can result from lack of clarity over test function can be found in Noss, Goldstein and Hoyles (1989) who document the situation arising in the United Kingdom when so-called graded tests were used both for instructional management and for certification.

The issue of fairness is fundamental to measurement validity, and is often misunderstood. Essentially, a fair test is one that is comparably valid for different groups, people, and settings. This requirement does not mean that all groups must get the same scores on a test, but that the scores should not under- or over-represent the knowledge or skills of a particular group. If one group has had very poor and disadvantaging educational opportunities in mathematics, and thus does not have a high level of knowledge and skills in mathematics, it is entirely appropriate that it should obtain lower scores on a test of mathematics than an advantaged, well-prepared group. In this case the test is fair: what is not fair is the prior unequal provision of educational opportunities.

An example of the existence of bias would be if culturally specific information were included in a test. Type of bias can be seen in a language proficiency test developed by the HSRC in the early 1990s. The text which forms the basis for numerous comprehension questions is full of such phrases as "catching a Blighty one" – i.e. receiving a wound which, while not fatal, is serious enough to have one invalided out and sent back to England (Blighty). Reliance on such a text can hardly be considered to be unbiased, as few children raised in Africa in the last years of the 20th Century could be expected to be familiar with such references.

A more complicated example of bias is illustrated in the following critical comment on the SAT:

> "Student performance on the test is influenced as much by the nature of household dinner-table conversation as it is by formal school instruction. That is, the vocabulary of households with a high socio-economic status is the vocabulary of the examination" (Adelman 1999:B4).

Clearly, this vocabulary will advantage middle-class candidates. The makers of the test argue, however, that in terms of the aim of the test - predicting performance in Higher Education - this so-called biasing factor is fully justified, in that the language of the test is based on the language predominantly used in Higher Education.

The issue of bias is a sensitive one, open to different interpretations. In addition to use of various statistical techniques such as Differential Item Function (DIF) analysis, trained review teams which check for evidence of bias can ensure that many of the difficulties are avoided.

3.4.2 Attribution of Cause

A crucial assessment issue that is frequently overlooked, particularly when a test has established a tradition of use, is attribution of cause. In the South African situation, for example, poor pass rates in most of the provinces are, rightly, the cause of much dismay and despondency. The fact that candidates have performed so poorly is, however, usually and again rightly, blamed on the poor preparation they have received for the examination, as well as on the examination itself. What is not commonly (if ever) highlighted as an issue – in public discourse at least – is whether it was fair or legitimate in the first place to have required poorly prepared candidates to write the examination, particularly when it is known that they had not covered the syllabuses on which the examinations were based.

This question of legitimacy is a complex issue, illustrated by the recent calls by the Congress of South African Students (COSAS) to be tested "... only as far as they had been taught in their syllabus" (COSAS president Lebogang Maile, quoted in *The Star*, September 12, 1999). Indeed, it is difficult not to sympathise with the plight of these candidates, whose teachers had completed less than half of the required syllabus in many subjects. This situation is captured in the following examiner's comment:

> I came across many scripts where students claimed that they met books for the first time in the examination room. As a result of this, a number of candidates handed in their answer books without anything written on them. Most of the candidates wrote little notes at the end of their answers to say that they did not read the books, they did not understand, they were not taught and that they read books that were not prescribed" (DoE 1998:20).

The American experience is interesting in this regard. The persistent gap in test scores between black and white candidates, for example, is a classic 'attribution of cause' issue. The cause of the poor performance is contested, however. Test development agencies cite poor educational preparation (e.g. The College Board 1999a&b, 1997, 1990), whereas minority community leaders lay the blame at the door of test bias (e.g. Garcia et al 1999, Ting & Robinson 1998).

It is important to note that attribution of cause is as much an educational issue as it is one of assessment. If disparities in educational provision exist, it is these that should be the primary target of increased efforts to increase test performance. However, it is also incumbent on the designers of tests to ensure that the tests are sensitive to such disparities, and that information about performance on tests is sensitively conveyed with due acknowledgement of the causes of poor performance.

3.4.3 Effectiveness of Treatment

The third of the criteria adopted by the Committee on Appropriate Test Use, effectiveness of treatment, refers to the consequences of performance on tests. If decisions are to be made about a candidate's future, it is essential that the test enable the candidate to demonstrate, fairly, what s/he is capable of. Cummins (1984) provides chilling evidence of the misdiagnoses of minority children in Canada who, on the basis of tests written in the medium of a language (English) in which they had not yet developed age appropriate cognitive academic language skills, were labelled 'educable retarded' and placed onto special education tracks. Chapter Eleven discusses the issue of consequential validity in some detail.

3.5 Conclusion

Chapter Three has highlighted several issues related to assessment which need to be addressed if the assessment procedures adopted for use in a particular context are not to subvert the educational processes that precede the assessment. Particular areas of concern elaborated in the chapter are 'fitness for purpose' issues, and the dangers consequent on test misuse, as well as issues surrounding test fairness. These issues are of particular importance in a context like South Africa, where educational opportunities are still grossly skewed and unequal.

In this connection, particularly, the relationship of tests to existing curricula was highlighted as a crucial area of concern. The relevance of this relationship was noted for the South African context, where an existing national school-leaving examination holds a prominent place in the embryonic quality assurance system in the country. This examination is a curriculum-aligned achievement assessment, and any proposed new assessment procedures at the school/Higher Education interface would need to forge strong links with it in order to avoid conflict and/or unnecessary duplication. As is argued in this study, however, total reliance on an achievement test, such as the school-leaving examination mentioned above, is not legitimate, or useful, in a context of unequal educational provision. Indeed, it can be argued that test fairness issues provide a large part of the impetus for the quest for additional selection mechanisms that go beyond simply reflecting that some candidates have had better educational opportunities than others. The need for additional, new assessment procedures that are sensitive to prior opportunities to learn is urgent.

The issues discussed in Chapter Three are fundamental to an understanding not only of some constraints, but also of some positive contributions that assessment could make to teaching and learning processes. Chapter Four provides a discussion of the role that assessment could play in the area of selection to Higher Education.

CHAPTER FOUR
ASSESSMENT AND SELECTION FOR HIGHER EDUCATION

4.1 Introduction

The previous two chapters have provided an overview of relevant and important general issues related to selection for Higher Education (Chapter Two), and to assessment (Chapter Three). In Chapter Four, the discussion focuses on the role that assessment procedures and approaches can play in selection to Higher Education. In other words, the chapter brings the two areas together.

4.2 Assessment and Selection for Higher Education

In many countries, selection for Higher Education is based primarily on the examination that takes place at the end of schooling. In many instances, this is the only external examination, although in some cases, students write an external examination at the end of an earlier stage (e.g. the "O"-level type examination in the United Kingdom, or the old Junior Certificate in South Africa). Some educationists have argued that, ideally, selection, monitoring and certification functions should be performed by different tests (e.g. Taylor & Vinjevold 1999, Resnick & Resnick 1992). According to Heyneman and Ransom (1992), this ideal is rarely achievable in developing countries such as South Africa, and so one examination has to perform a number of functions. Reasons for this include the obvious one of financial resources. However, just as crucially, the scarcity of suitably qualified local assessment specialists (Altink & Thijs 1984) makes it difficult to develop multiple

testing systems to meet these needs. Instead, one test or set of tests at the end of schooling is used to select students for further educational opportunities as well as to certify mastery at the end of schooling and to signal readiness for employment. In addition, in many developing countries, the examination is often the only means by which quality can be assured.

In the virtual absence of any effective set of procedures for monitoring the quality of schooling, it seems inevitable that the external school-leaving examination will be called on to serve as a rudimentary quality control mechanism. Unfortunately, the lack of capacity which has made it difficult to establish effective school-based quality promotion systems can also be seen in the very limited and often counter-productive ways in which the information that is available from the examination is used. Examples of problems in this regard in the South African context are contained in the Report of the Ministerial Committee on the Senior Certificate Examination (DoE 1998), which details, *inter alia*, the dysfunctional relationships existing between the examination and curriculum sections within many provincial education departments. Other examples are found in Taylor and Vinjevold (1999) and the Curriculum 2005 Review Committee Report (DoE 2000).

The multiple uses of external school-leaving examinations, and in particular their gate-keeping function, means that such examinations have a very high public profile. While this high profile can have positive benefits, in that learners are more likely to be motivated to stay at school and to prepare for the examination, it can mean that innovations and changes are difficult to introduce.

A further characteristic of external examinations in developing countries relates to the serious administrative and security problems, resulting from high levels of corruption amongst officials, and from marginal rural infrastructure such as poor roads and systems of communication. These problems, which impact on the integrity of the examination, contribute to the unease, which surrounds the results of the examination, and the tendency on the part of admissions officers at Higher Education institutions to seek additional indicators of merit.

There are very large disparities in educational provision and quality within the schooling systems of developing countries, and these are reflected in the results obtained by candidates. This is a

characteristic of all systems, it is acknowledged. However, in developing countries the situation is exacerbated by greater gaps between rich and poor; the fact that the language of the examination is frequently the second or further language of the majority of the candidates; and, crucially, by the fact that for many candidates in developing countries, achievement is more directly linked to school quality (Muller & Roberts 2000, Heyneman & Ransom 1992) than in so-called developed countries. This last effect arises because many schoolchildren in Africa come from homes where the parents are illiterate – in their home language as well as in the language of the school - and where public libraries and other educational resources are virtually non-existent. In these cases, the school bears a far greater share of the responsibility for educating children than its counterparts do in richer societies, and the quality of the school accounts for more of the variance in outcomes.

It follows, then, that in developing countries, where the external examination plays a very strong role in admissions decisions, because of the lack of viable alternative indicators of merit, the problem of selection fairness raises its head most acutely. Since, as has been argued above, educational opportunities are more directly linked to school quality in poor nations than in rich ones, it is unfortunate – while understandable - that further opportunities in poorer nations are so firmly linked to school performance as demonstrated on school-leaving examinations. The need for additional sources of information is clear, and it is ironic that it is in richer countries, such as the United States where the need is not so great, that such sources are currently more readily available.

4.3 Research into Higher Education Selection Practices

To the layperson, the lack of clarity over what 'works' in admissions must seem inexplicable. In common sense terms, it seems a straightforward matter to conduct research into the issue, simply by admitting a range of students, then seeing who succeeds and translating important distinguishing features of success into admissions criteria. Some of the major difficulties and complexities in admissions related research, which make such a seemingly straightforward undertaking very difficult, are outlined below.

4.3.1 General Problems

The problem of truncated samples arises when only a subset of the original test population can be used in the validation process. For admission projects, this problem will usually mean that only the students who performed well in the original assessment procedure will be available for validation purposes, as they are the only ones selected on the basis of the procedure and hence admitted – i.e., unsuccessful students are rarely part of the validation sample. Tests used for placement purposes – that is, to place students onto a particular course of study - suffer from the same problem in respect of validation. This arises because only students whose scores indicate a particular phenomenon – for example, they are below a particular threshold - are likely to be placed on remedial courses. The problem is compounded when tests are used for both purposes – that is, for both admissions and placement, as is frequently the case with admissions projects. Attempts to validate either the courses onto which students have been placed, or the placement mechanism, by using the scores of the placement/admissions tests will almost inevitably yield only inconclusive results. As Linn (1983) suggests, a highly selected sample from a specified upper (or lower) range of a predictor frequently yields an unrealistically pessimistic view of predictive validity. Put differently, truncated samples tend to depress the correlation coefficient, because the vindicating evidence from the eliminated group is absent, for ethical or financial reasons (or both).

Small sample size is a perennial problem in selection studies, too. One of the difficulties is that enlarging the sample size by incorporating groups that are not strictly comparable to the sample group often leads to other problems. For example, an investigation into the predictive validity of a selection mechanism for Chemical Engineering should be limited only to students in that programme of study. The number in the programme is unlikely to be large. However, enlarging the sample size by including other Engineering programmes such as Mechanical or Electrical will undoubtedly limit the strength of the finding, as the different programmes have different pass rates, different courses, and different areas of challenge for students, and these differences impact upon predictive validity associated with the original selection process.

Studies into selection practices are also plagued by difficulties over defining the criterion of success. Various studies have defined success differently, and this has tended to make comparisons and generalisations very difficult. Examples of options in this regard include the following:

- passing a stipulated proportion of the courses taken in first-year;
- through-put to graduation, usually in a specified maximal period; and
- quality of performance – for example, some studies weight performance in terms of whether students achieve 50%, 75%, etc.

Decisions about these options can make a marked difference to the way in which selection procedures are developed. For example, if the aim of the selection is to identify only those candidates who will obtain first-class passes, the tests can focus on the upper levels of difficulty, and not bother with obtaining ranking information about all candidates. If the aim is to distinguish between those who will get above 50% and those who will not, the important aim will be to develop items that distinguish between candidates at this level.

Another source of difficulty can be seen in investigations into selection procedures that attempt to identify potential abilities of the applicants, and not to rely wholly on school results. Such studies run into serious difficulties in establishing realistic validity indices. Altink and Thijs (1984:75) point out that the concept of potential ability poses a problem for validation, as

> "[T]he predictors used have a different conceptual character. Achievement predictors are less remote from the criterion and may, as such, show more validity than ability predictors, which are derived on the basis of hypothetical relations between predictor and criterion variables."

For example, the starting point of most South African first-year Mathematics university courses is closely aligned to the SC Mathematics HG syllabus. It follows that performance on the Mathematics HG examination, an achievement test, ought to be a reasonable predictor of end of first-year university Mathematics courses. However, a test of ability, or of potential, will not align so closely to the syllabus of any one course. Moreover, since the attributes tapped by ability tests take time to develop, short-term criterion measures will advantage achievement tests in terms of

predictive validity. For this reason, comparisons of the predictive validity of the two types of test require long-term criterion measures, such as graduation rates. The absence of long-term criteria is one of the major criticisms of many of the ETS studies of the predictive utility of the SAT, which take first-year academic performance as the criterion (Garcia et al 1999, Crouse & Trusheim 1988). Zaaiman's (1998) study of selection procedures in a South African context (see 4.3.2 below) is also vulnerable to this criticism.

However, time between predictor and criterion is another source of difficulty for selection studies. Thus the use of long-term criterion measures, as recommended above, while addressing one source of difficulty, introduces another in that the number of variables affecting performance can increase with time, and each of these variables may serve to render the predictor-criterion linkage less discernable by inducing variations of their own.

The impact of educational interventions, a further source of difficulty for selection studies, makes the interpretation of selection studies tied to some form of educational consequence extremely complex. Attempting to ensure that needs identified by selection tests are met through appropriate educational provision sets up an inevitable tension, because if the students' needs are indeed met, they will be less likely to fail and the selection test, which showed them to be at risk, will appear to be less predictive. This kind of problem is described by Altink and Thijs (1984) as "the incongruity between 'maximum prediction' and 'equity'". That is, if one wanted to demonstrate the strength of a selection instrument, one should admit all students and let them sink or swim without intervention. However, many selection programmes also act as placement mechanisms, and this use often means that the students identified as most at risk obtain the most assistance, and can actually do better at the institution than students who appeared stronger on the original test but received less assistance. This consequence makes selection studies tied to educational interventions very complex, and inevitably compromises the predictive study[20]. In the project described below (4.3.2) for example, the selection tests had an iterative relationship with the actual courses comprising the pre-entry year. Thus the problems which manifested themselves in the tests of the selected

students were targeted in the courses, and, likewise, problems found to be significant in the learning progress of selected students were included in subsequent versions of the tests. Clearly, this interaction complicates any interpretation of relationships between the predictors and the criterion in a given year. If there is a high correlation, for example, do we infer that the courses failed substantially to meet the identified learning needs of the students? Conversely, if there is a low correlation, do we infer that the predictor was at fault and that the problems manifested in the students' test scripts were not important?

The sources of difficulty outlined above go some way to clarifying why research into selection practices is difficult both to undertake and to interpret. The selection projects described in the following section illustrate many of the difficulties.

4.3.2 Selection Research Projects

The projects sketched below represent studies of selection practices. The intent of this section is not to provide a comprehensive review of international selection practices in various regions of the world, but rather to highlight some of the challenges, findings and constraints of research into selection practices. Comprehensive reviews of the ways in which various countries conduct their admissions and selection procedures can be found in, for example, Zaaiman (1998), and Ball and Eggins (1989).

In the USA, studies on selection practices abound. It is difficult to make generalisations about admissions and selection procedures, as the United States has an extremely diverse and large Higher Education system. As local control over all phases of schooling is jealously guarded, there is no system of state or national school-leaving examinations such as exists in South Africa and the United Kingdom, for example. One consequence of this diversity is that Higher Education institutions have turned to external, standardised tests for the provision of comparable information on their applicants[21]. The most widely used tests are the SAT tests, taken by more than one

[20] As is demonstrated in Chapter Ten, however, the AARP study has avoided this 'incongruity' as students who were not selected on the basis of their AARP test scores are included in the study, as the institution did not use the AARP scores to exclude students.

[21] A different approach to establishing comparability between scores can be found in the Queensland Core Skills Test, which is used in combination with school-based information. This is described in Chapter Five

million students in each high school graduating cohort, and the American College Testing (ACT) tests. The SAT tests are of two types: the SAT1: Verbal Reasoning Test (usually known simply as the SAT), and the SAT II: Subject Tests. The latter are achievement tests, which tap content knowledge in school subject areas. They are taken by far fewer students than take the SAT1 tests, but are found useful by admissions officers when used "... in conjunction with what teachers write about a student" (Hernandez 1997:53) to establish a standard that can be applied nationally. The ACT and SAT II tests differ in that the SAT II tests are subject tests, while the ACT assessment is based on four sub-scales: English, Mathematics, Reading, and Science reasoning, all of which tap core competencies which form part of school curricula. The ACT assessments thus can be said to occupy the middle ground between two kinds of SAT tests, with the SAT1 focusing on developed verbal and mathematical reasoning skills, and the ACT focusing on achievement related to high school curricula.

With reference to the SAT1, the most widely used of the tests, a publication of the College Board states:

> "Literally thousands of validity studies have been conducted for institutions throughout the country for more than a half-century. The evidence is clear: the SAT works and it works very well in many different circumstances. However, just as with average scores, there are differences in how it works for different groups of students, for different types of educational programs, and for different institutions" (The College Board 1997:4).

Studies investigating these differences are legion. In essence, they confirm that, when used in combination with other sources of information such as high school grade point average and place in class, they improve the correlation with college grade point average to about 0.55 (op cit). Similar findings are reported about the ACT tests (The College Board, 1999b). Even with databases as large as those available for research into the SAT and ACT tests, however, difficulties are experienced in interpreting the results. Truncated samples are found, as only self-selected groups of students write the tests which, on the whole, are required only by elite colleges, and then only some of these students are selected for admission.

One of the most difficult challenges for admissions officers seeking to constitute a racially diverse student body arises from the recent rolling back of affirmative action admissions policies in the USA. Proposition 209, as it is known, ended race-based admissions policies in California's public universities. As Garcia et al (1999:5) suggest, "[T]he bans there on utilising race, ethnicity and gender as an admission criterion have sent women and underrepresented students tumbling off the playing fields of selective public universities". A recent issue of the *Chronicle of Higher Education* devoted three full pages to an article entitled: *"A Top University Wonders Why It Has No Black Freshmen"* (28 April, 2000). While the thrust of the article was on the reasons for black students not opting to apply to that particular institution, the article does highlight the difficulties experienced by institutions which are no longer able to use test scores in ways that reflect differentials in group performance.

One of the most comprehensive studies into the effects of race-sensitive admissions criteria is that of Bowen and Bok (1998). Using a vast database (the records of some 80,000 students at 28 highly selective colleges, starting from 1951), they investigated two major issues: the success rates of students who had been admitted as a result of some kind of affirmative action programme; and the effect on the numbers of minority students if affirmative action were abandoned. In answer to the first question, they demonstrated that the graduation rate of the sampled black undergraduates, while behind the white rate (79% versus 94%), far exceeded the 32% rate for black students nationally. Clearly, the black students in the study constitute an academic elite, even if they gained admission on the basis of affirmative action rather than direct competition, and thus the graduation rate of 79% is not strictly comparable with that of other black students. The contribution to society of such a high number of successful, well-educated black graduates is inestimable, however, and begs an answer to the second area of investigation – how many of these graduates would have been at these institutions had affirmative action not been applied? According to Bowen and Bok's analysis, black enrolments would fall from the current levels of 7.1% to 2.1%, and to 1.6% at the most selective colleges. In Law and Medicine, the number of black students would sink to less than 1% at these colleges. It is thus clear that the majority of the graduating black students has

been admitted on affirmative action procedures, and that the introduction of race-blind admissions procedures would sharply curtail the numbers of black graduates from selective colleges.

In the United Kingdom, the formal entry requirements to most degree courses are two A-levels at grade E or above, or their equivalent. Most institutions require higher grades, however, and several combine school results with interview protocols. An interesting feature of the UK admissions scene is that, overwhelmingly, access routes to Higher Education (other than the traditional A-level route) apply to adults and not to school-leavers. In general, these routes take the form of specially designed access courses (Wagner 1989, Woodrow 1988, 1986). Woodrow (1988, cited in Wagner 1989) suggests three characteristics of access courses[22]: they are targeted at under-represented groups; they are delivered collaboratively by the receiving and access course institutions; and they are part of the higher education studies, and not simply preparation for them. In some cases they are linked to a particular course, and in many cases some form of credit is given.

It can be argued, then, that in the UK the activity around admissions and access has tended to concentrate less on selection procedures, and more on the development of admission routes involving instruction. However, recent reports indicate that the low numbers of students from state schools gaining access to highly selective institutions - such as Oxford - are a cause for concern, and that some form of tests of 'core intellectual skills' could be considered for introduction by individual institutions in the future (BBC News/UK Systems 1999).

Ajayi et al (1996), in a comprehensive review of Higher Education in Africa, report on the difficulties faced by Higher Education in general, such as the struggles to overcome the effects of colonialism, decreasing resources, and increasing political control. They identify specific challenges associated with access as follows: little institutional autonomy in terms of selection of the student body; discrimination against women and certain ethnic groups; and large-scale reliance on school results although great inequities exist in school quality and socio-economic status.

[22] An example of an access course (Zaaiman 1998) in the South African context is described below.

Little systematic research has been conducted into selection procedures and practices in Sub-Saharan Africa generally. As is the case with the United Kingdom, the research that has been conducted has almost entirely involved the development and evaluation of access routes, where selection and curriculum development and delivery are inextricably linked.

A small number of studies relating to selection for Higher Education in Sub-Saharan Africa north of South Africa has been conducted in collaboration with, and supported by, the Vrije Universiteit Amsterdam (VUA). These include studies by Drenth et al (1983) in Tanzania in collaboration with the University of Dar-es-Salaam, and in Kenya in collaboration with the University of Nairobi (Kenyatta College). These studies both concluded that use of ability and aptitude tests, by reducing reliance on schooling systems known to provide inequitable opportunities, would increase fairness and equity in selection in those countries.

In addition, Altink and Thijs (1984) report on the selection procedures developed and used for pre-entry science courses in the Boleswa countries, viz. Botswana, Lesotho, and Swaziland. The aim of these programmes is primarily to increase the numbers of students in science-based tertiary programmes, with the ultimate aim of improving science and mathematics education in those countries. In keeping with their findings in regard to the VUA studies mentioned above, they note that "[S]chool records and school examination results are insufficient indicators of the potential abilities of the applicants, because of the variation in quality of the schools" (op cit:75). Because of this inadequacy, ability, aptitude and achievement tests were used in their selection procedures.

According to Zaaiman (1998), the ability tests aimed to tap numerical, verbal and spatial abilities, as these were "presumed to facilitate the process of learning" (Zaaiman op cit:72). The aptitude tests were designed to assess core skills such as problem-solving, application, extrapolation, insight, applied to concrete problems embedded in science and mathematics contexts. The aptitude tests deliberately minimised content knowledge. The achievement tests used were the final school examinations (Cambridge Overseas School Certificate examinations). As might have been expected, the best predictors of performance on the pre-entry science courses were the achievement tests. However, the research did not continue beyond the pre-entry year, and so it is

not known what the predictive power of each type of test would have been at the time of graduation. The work of the VUA projects was taken further in the University Foundation Year (UNIFY) project in South Africa, which is discussed below.

Research into selection practices in South Africa is characterised, in the main, by reliance on small numbers and a high degree of context-specificity (Griesel 1999, Zaaiman 1998, Yeld & Haeck 1997). Most admissions-related research projects, as is suggested above, are further complicated by being attached to curriculum development initiatives of some kind, such as the College of Science, the Teach-Test-Teach (TTT) and the UNIFY projects discussed below. In effect, this attachment shapes the research as an investigation into effective placement and curriculum development activities as much as selection procedures. This coupling of admissions to placement procedures is, as has been argued above, a necessary consequence of the great disparities in, and generally poor quality of, educational provision in schooling. However, the coupling inevitably limits the generalisability of the findings to the programmes for which placement was designed. In other words, tests, when used for diagnostic purposes and as a basis for curriculum design, become difficult to research as admissions instruments. This point is illustrated below in the discussion on the UNIFY project.

The major selection-related projects and initiatives in South Africa are: the UNIFY programme in Mathematics and Science at the University of the North (UNIN); the College of Science at the University of the Witwatersrand; the TTT project at the University of Natal; the UWC Lottery system; and the AARP project at UCT.

The aims of the UNIFY programme are to increase the numbers and improve the academic quality of the students entering and studying in the UNIN faculties of Mathematics and Natural Sciences, Health Sciences and Agriculture. To achieve these aims, the UNIFY project sets out to prepare students for entry to these faculties, by giving them a strong foundation year in relevant discipline and skill areas.

The achievements of the project can be listed as follows (Zaaiman 1998, Zaaiman 1996, Griesel 1999):

- The selection tests for the UNIFY years 1994-1996 consistently show positive significant predictive validity for the UNIFY final results. The correlations between UNIFY final results and individual selection tests range from 0.28 to 0.57, and probably underestimate the position for the whole applicant group. The underestimation arises because the calculations were of necessity performed on a highly selected group, a truncated sample, as discussed above.
- The selection tests for the UNIFY programme do not predict the performance of ex-UNIFY students in their first year at UNIN (that is, for the year following the UNIFY year). Zaaiman (op cit) gives small sample size as a possible reason for this lack of predictive validity. Other reasons, which apply to some extent to all cohort studies, include such factors as the length of time between predictor and criterion performances, personal contexts - such as financial situation, accommodation, personality, motivation - and differences in standards and instructional methods in academic courses.
- Notwithstanding the above, Zaaiman's study (Zaaiman 1998) shows that UNIFY final results (as opposed to the UNIFY selection tests) are a good predictor for BSc first year performance. Indirectly, therefore, since the UNIFY selection tests are good predictors of UNIFY final performance, and UNIFY final results are a good predictor for first-year performance, it can be argued that the UNIFY selection tests can be used to 'produce' successful UNIN first-year BSc students. Unfortunately, small sample sizes precluded the undertaking of a rigorous study of the predictive validity of the UNIFY selection tests for the first-year BSc performance of non-UNIFY students. Such an analysis would be useful as a means of investigating the contribution of the UNIFY year itself, as opposed to the selection tests.
- In terms of the project more broadly, UNIFY students make up a significant proportion of first-year students in the target faculties. That is, the projected increased though-put into UNIN first-year science courses has occurred.
- The ex-UNIFY students consistently outperform their non-ex-UNIFY peers at the first-year level. As Griesel points out, however, it is "... widely established in this model of intervention

(an add-on preparatory year) that student performance levels typically decline sharply after their first year in mainstream study" (Griesel 1999:47). The tracer studies conducted thus far on UNIFY student performance (Zaaiman 1996 and 1998) track student performance only until the end of the first academic year, and the absence of post-first-year data and analysis is undoubtedly a limitation of the findings.

The Zaaiman study did not include considerations of cost or sustainability. It does demonstrate, however, that poorly prepared students can succeed given appropriate intervention. As a study into selection its generalisability is limited because of its strong and iterative link with the UNIFY curriculum.

The College of Science at the University of the Witwatersrand, like the UNIFY project, is an integrated extended curriculum project. Students are selected on the basis of a number of tests in addition to a biographical questionnaire. These are: a multiple-choice format Science aptitude test, a logical sequencing test, a multiple-choice Mathematics test, a Physical Science test, and a spatial ability test. The means of the five tests are calculated, and students are selected who are above the mean on at least three tests. Students who are above the mean on two tests might be selected for interview (M. Rollnick, email communication, 14 Feb 2001). The tests have not been updated for several years, and other than an early tracer study into the progress of the students, which showed promising results in line with similar programmes at other institutions, no further selection-focused research has been undertaken.

The TTT project at the University of Natal was responsible for designing three selection programmes for students aiming to study Social Science, in 1988, 1989 and 1990, and the subsequent application of the selection model to an entrance examination in 1991 and 1992. Like the UNIFY and College of Science initiatives, the TTT programme "... is involved in educational intervention both prior to selection and in selected students' first year of study" (Griesel 1992:59). The entrance examination takes the form of a series of foundational textbooks which are mediated through the provision of learning guides. Students write a number of assignments, followed by an open-book examination. The focus in the process is on the "extent to which students have

benefited from the learning opportunities provided prior to the examination" (op cit:62). However, there appears to be little or no empirical basis for measuring this improvement, as students are not assessed at the beginning of the process (Griesel, personal communication, 1998), and so it is not possible to assess the success of the initiative in terms of selection: it needs to be assessed on its own terms, namely, as an educational intervention.

During the late 1980s, UWC experimented with a lottery system for admission to its Arts Faculty (Herman 1995). This initiative is discussed in Chapter Two (section 2.4.3) in some detail, and the details are not repeated here. Recent developments at the institution include the abandonment of the lottery system in favour of quite extensive use of the 'Senate's Discretion route' (Tahir Wood, personal communication, December 2000). This route entails the establishment of Senate-approved admissions procedures. The reasons for the demise of the lottery scheme are not publicly available, but it appears that unacceptably high failure rates forced a review of admissions procedures, and that institutional research into the predictive utility of the Senior Certificate examination pointed to a stronger than anticipated relationship between even very poor results, and performance at UWC.

The AARP Project results are discussed in detail in Chapters Eight to Twelve. In very general terms, the AARP annual reports consistently make the point that good performance in either the AARP assessments or the SC (or both) is a reasonable predictor of success at UCT. Indications are, however, that for students from educationally disadvantaged backgrounds[23], students who achieve good results on the AARP tests will have lower exclusion and higher graduation rates than students who do poorly on the tests, irrespective of their Senior Certificate scores. From a research point of view, this project holds potential, as it tests sizeable numbers of candidates, and recommends students for all programmes at the university. It is thus somewhat independent of specific instructional effects. In addition, as it has not in the past been used as a barrier to access – that is, poor performance on the Project assessments has not been used to reject applicants who would otherwise have been accepted – the sample is not as truncated as is often the case with

[23] At this stage longitudinal data are not available on the performance of students who are not educationally disadvantaged, as such students only became eligible to write the Project tests relatively recently.

selection studies. Therefore, as can be seen in Chapter Ten, it is possible to compare the subsequent academic performance of students who had done well enough on the Project tests to have been recommended for admission on that basis, with that of students who had done poorly on the Project tests, as the latter group would have gained access to the institution on the basis of their School Leaving results.

It can be seen from the discussion above that although several small-scale projects have been undertaken in South Africa, they have tended to be tied to curriculum development initiatives, and thus are difficult to evaluate in terms of selection. The AARP project at UCT is an exception to this, and results to date are promising, although the necessary longitudinal data are only available in relation to the performance of educationally disadvantaged (ex-DET) students.

4.4 Static and Dynamic Approaches to Assessment

The discussion below focuses on non-curriculum aligned selection tests in the context of selection for Higher Education. Achievement tests, as has been argued in Chapter Three, are based on a preceding course of instruction and are not as profoundly dependent or reflective of an understanding of human cognition as are non-curriculum aligned tests. That is to say, achievement tests aim to assess the extent to which an individual has mastered a body of knowledge that has been taught, or has been assumed to have been taught. The extent to which this material has been mastered, it is believed, will serve as a useful indicator of future academic performance. Non-curriculum aligned selection tests, in contrast, set out to investigate the extent to which students have developed the kinds of cross-curriculum, core skills and abilities that are believed to be important in the majority of learning situations that will be encountered in Higher Education learning environments.

It has been argued repeatedly in this study that the kinds of information that can be gained from end of school achievement tests, such as the Senior Certificate in South Africa, are of great value in selection processes, and that their importance should not be undermined. However, the educational inequities that exist in the country are such that reliance only on achievement tests will

severely disadvantage those who are not able to compete on an equal basis. The discussion in this section subjects non-curriculum aligned testing approaches, which need to be demonstrably capable of adding valuable information to the kinds of information yielded by achievement tests, to scrutiny. As a starting point, two major approaches are analysed.

4.4.1 Static Assessment

In this approach, the harmful effects of bad instruction are ignored or downplayed. The assumption is that individuals will be able to perform satisfactorily on a test provided that it is not linked to any particular curriculum. This view leads to the search for a test that will be independent of prior learning - often described as being `content-free'. For example, in the 1960s, potential and existing test users (usually colleges) were encouraged to think of the SAT as "...an aptitude test that predicted success in college and saw through the veneer of poor preparation and social and economic circumstances" (Crouse & Trusheim 1988:36). The test is, according to its developers, not a measure of 'total intelligence', but "... measures only verbal and mathematical reasoning skills in order to predict the college grades of freshmen" (The College Board 1999a:3).

Evidence is mounting, however, on the crucial role played by domain knowledge in the development of expertise. This role is discussed in Chapter Five, and is aptly summed up by Resnick and Resnick (1992:41) as follows:

> "... recent cognitive research teaches us to be highly respectful of knowledge .... Study after study shows that people who know more about a topic reason more profoundly about that topic than people who know little about it ... . Education requires an intimate linking of thinking processes with important knowledge content."

The implication here is that people learn to 'reason more profoundly' – to make connections, to ask questions, to elaborate – while engaged in learning about something, and that not being exposed to these opportunities which are maximised through meaningful conceptual development, adversely affects individuals' cognitive development.

Research conducted by, amongst others, Greeno and Simon (1984), Siegler (1983) and Chi (1978), reveals that "[C]ontent in terms of knowledge base ... [is] ... an important distinguishing factor between good and poor learners" (Lidz 1987b:464). Downplaying the importance of content

can thus blur the distinction between effective and ineffective learners, whose detection, in selection testing, is obviously a fundamental aim.

Indeed, the continuing controversies around the use of (for example) the SAT's and national assessment programmes in the United States (Lemann 1999a&b, Gipps & Murphy 1994, Langer, Applebee, Mullis & Foerthsch 1990, Crouse & Trusheim 1988), suggest that it is not, in fact, possible to develop `content-free' tests. Increasingly, it is accepted that tests developed within this paradigm, including so-called intelligence tests, do rely on prior learning and inevitably privilege certain groups. This is acknowledged as follows:

> "Low average scores for some groups reveal problems in education and society ... . When faced with unpleasant news, it is human nature to seek a scapegoat or kill the messenger. Those who claim that score differences are the result of test bias must ignore ... the unequal educational opportunities that reporters, educators, and researchers have documented for decades in barrios, inner cities, and isolated rural areas" (The College Board 1999a:3).

It is interesting to note that this rather defensive comment by the College Board refers to the same test as that touted as 'seeing through the veneer of poor preparation' as quoted above.

It is argued above that so-called 'content-free' tests cannot fulfil the claims made on their behalf. That is, they cannot, simply by ensuring that they are not linked to any particular curriculum, in themselves 'level the playing fields' for all candidates; nor can they provide a satisfactorily comprehensive picture of the status of learners' cognitive achievements. Indeed, it is by now generally accepted amongst educationists and psychologists that " ... conventional intelligence tests can result in an underestimation of ... real intellectual potential" (Hamers & Resing 1993:27). This underestimation is particularly likely to occur when the candidates being tested have had inadequate educational or social opportunities to develop their knowledge and skills in comparison with their more privileged peers. It follows that the legitimacy of 'content-free' tests, in the unlikely event that it were possible to develop such tests, rests largely on the homogeneity of the educational experiences which precede the testing, and on the validity and reliability of the testing procedures and instruments.

Dissatisfaction with traditional approaches to assessment - in particular the psychometric tradition exemplified in IQ tests - is summed up by Bransford, Declos, Vye, Burns and Hasselbring (1987), who categorise limitations in the following three broad ways. First, traditional ways of testing focus only on the products of learning, and not on the learning processes that result in these products. This is clearly a major issue in contexts of grossly unequal learning opportunities such as those found in South Africa, as it leads to inferences on ability being made on the basis of demonstrated performance only, irrespective of the opportunities to learn that have preceded the elicitation of performance. In terms of test fairness, this is clearly unacceptable. Second, since traditional approaches do not provide meaningful insights about processes of learning, they do not offer any pointers for effective intervention. Third, traditional approaches to assessment do not tap responsiveness to instruction and thus the limited amount of information that is gained has nothing to say about how effectively an individual might learn if given appropriate opportunities.

This framework for selection practice and related research is, as can be seen, firmly located in a static notion of intelligence as a phenomenon which simply has to be pinned down and measured.

The second approach to testing, however, treats intelligence as a dynamic phenomenon, always under construction, always shaping and being shaped by the environment in which an individual finds her/himself[24]. It is aligned to the cognitive and situative perspectives on knowing and learning discussed in Chapters Five and Six below. Selection practices in this framework aim to assess the ability of an individual to respond to educational intervention, and thus rest on the assumption that appropriately designed mediation can rectify the influence of poor prior learning experiences on test performance. They therefore require the incorporation of a teaching/learning element into the selection procedure. In addition, they assume that, in one way or another, the selection process and the receiving institution will take seriously the need to " ... unravel the nature of the conflict between different learning histories and what this demands of the learning/teaching situation"

[24] This assertion is not to suggest that because 'intelligence' is always under construction, it cannot (in principle at least) be assessed at any stage. Rather, what is being asserted is the dynamic nature of intelligence, and the impossibility of ignoring the role of past educational experiences in constructing current or future trajectories.

(Craig 1991:137)[25]. Doing so, it is argued, will facilitate and promote the coupling of selection to appropriate educational responses post-selection. The implications of this last point, and the consequences of admitting talented but educationally under-prepared students to courses without proper support, should be taken seriously. As Sewell (1987:441) cautions, "... it should not be assumed that the high potential ..., determined by the degree of modifiability considered feasible, will be necessarily translated into school achievement".

The distinction between achievement and potential in this context is somewhat confusing, and can perhaps best be explained by drawing a distinction between – by way of example - passing Chemistry 1, and having produced indications of being able to pass Chemistry 1. Until the individual actually succeeds in passing Chemistry 1, this latter kind of accomplishment remains as potential – as the promise of something still to be demonstrated. This notion of potential is far softer than that referred to by Miller (1992:154), who questions the very notion of `potential' (in terms of testing) as follows: "What is appealing about the term `potential', as a name for an invisible target, is that it refers to the absence of something by affirming its presence." His argument is that in order for something to be assessed, that something must exist: a test must elicit a performance of some kind, and that performance is an indication of an individual's ability to produce the performance, not of potential to do so. His concern is echoed by Ryan (1972:42) who, referring to the notion of innate ability or potential, points out that "... the notion of potential ability both as something abstracted from all interactions with the environment and at the same time as something measurable in a person's behaviour simply does not make sense".

Given these difficulties, rather than suggesting the development of tests of 'potential', it seems more legitimate to aim for the development of a testing approach which aims to go beyond simply assessing what students know or have been taught how to do. It is clear that tests based on a syllabus will serve to perpetuate inequities unless prior opportunities have been roughly equivalent for all candidates. Since this is not the case in South Africa, it is clearly not acceptable to base

[25] The Russian psychologist Leont'ev phrased the challenge as follows: "American researchers are constantly seeking to discover how the child came to be what it is: we in the USSR are striving to discover not how the child came to be what it is, but how it can become what it not yet is" (cited in Hamers & Resing 1993:45).

selection decisions solely on achievement tests. It is further accepted that the idea that tests can of themselves, in some miraculous way, circumvent and/or negate poor learning opportunities has neither a convincing theoretical nor empirical basis. Thus tests which claim to test 'potential' simply by being non-curriculum aligned, and are by virtue of that not achievement tests, do not by that means succeed in removing the inequities arising from unequal prior educational experiences. The cause of poor performance on the part of some candidates might not, it is true, be directly attributable to a preceding course of instruction, but it is nevertheless traceable to unequal and disadvantaging educational experiences.

In brief, then, an approach is needed that is unlike that of traditional tests which are based on a syllabus or programme of work which has either been - or is assumed to have been - covered, and also unlike that of so-called 'content-free' psychometric tests of ability such as the SATs in the USA. Rather, tests that incorporate the kind of content and tasks that will both encourage and reveal concept and skill development, are envisaged. In other words, it is argued that the essentially chimerical nature of 'potential' can best be confronted though attempts to assess candidates' abilities to respond to appropriate educational processes/interventions. The challenge, if this is accepted, is to provide these interventions within the selection process, and, as was argued above, to ensure that appropriate educational responses follow the selection process. Specifically, the selection challenge becomes the construction of selection instruments and procedures that do not in effect 'predict the past' (Miller 1992), but are based on core academic skills as well as tasks that aim in part to teach students how to perform these skills.

A test within this paradigm would entail the provision of opportunities for action on the part of the learner, action in the sense of "... any kind of activity, mental or physical, that changes the way a situation (task) is experienced by the person who produces the action " (Miller 1989:156). It was argued above that "... those with the potential for university studies will show themselves when given the opportunities to do so" (Craig 1991:142), and the clear assumption is that traditional approaches to testing do not provide these opportunities, whether or not they are aligned to a

curriculum. Tests within this paradigm would thus conform to the notion that "... change is the *sine qua non* of assessment ..." (Sewell 1987:439).

Moreover, it should be noted that the development of this approach is recommended as an additional means of eliciting information about what students know and can do, and not as an alternative. This choice for more information arises because the purpose of the study is to enhance selection procedures in a situation in which, given historical and continuing inequalities in educational provision, no one measure of cognitive achievement is likely to be able to 'deliver' a fair and effective assessment, for all applicants, either in terms of what students know and can do, or in terms of what they have been taught.

4.4.2 Dynamic Assessment

Approaches to assessment which fall within the paradigm of dynamic assessment have been called, variously, the 'learning potential assessment approach', the 'learning ability assessment approach', and the 'dynamic assessment approach'. In this study, the term 'dynamic assessment' is used, primarily to avoid the confusion that can accompany the use of 'potential' which, as has been argued above, is not supported as a realistic goal of testing for selection.

In general, this approach to assessment is used in situations where for some reason it is believed that candidates will not be fairly or usefully assessed with more traditional approaches. Usually, the reasons for this belief are that the candidates will have been educationally disadvantaged by any of a number of factors. Dynamic assessment is commonly conducted within a 'test-teach-test' design framework: that is, an initial test (with no direct preparation) is followed by an educational intervention of some kind, and then by a second test. It is thus not so much an assessment of potential, which, particularly in its early days, it was sometimes called, but a test of ability to benefit from instruction.

The theoretical cornerstones of the dynamic assessment movement were laid by L.S. Vygotsky, a Soviet psychologist working at a time when the Soviet Communist Party had banned the use of intelligence tests (Guthke 1993). This ban essentially promoted the development of observational

techniques for assessing 'intelligent behaviour', which emphasised the interactive, social nature of human intelligence. Vygotsky distinguished two levels of intellectual functioning. One, called the Zone of Actual Development (ZAD), signified the level measured by most traditional intelligence tests, the level an individual can attain by her/himself, with no support. The other, called the Zone of Proximal Development (ZPD), signified the level reached after a training phase. As Vygotsky defines it, the ZPD is "... the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (Vygotsky 1978:86).

Extrapolating from this, it can be argued that an individual who has had the good fortune to attend an excellent school will have only a very small difference between her ZAD and ZPD on tasks resembling those acquired during formal schooling. An educationally disadvantaged individual, on the other hand, will have a far larger gap between the two. It is on these grounds that it is argued that educationally disadvantaged learners are further disadvantaged by assessment procedures which explore the ZAD but make no attempt to tap the ZPD, whereas educationally advantaged learners can be considered to be relatively fairly assessed on their ZAD and ZPD - through the use of traditional achievement tests, for example.

This view is strongly supported by research conducted by Babad and Budoff (1974) and Budoff and Friedman (1964) on the training effect of problem-solving strategies. Candidates were characterised as 'gainers' or 'non-gainers' on the basis of pre-and post-test scores, and it was found that 'gainers' tended to be children from low socioeconomic backgrounds. Further research evidence can be found in the work of Sewell (1979) who, like Budoff and colleagues, found high validity for middle-class children on traditional intelligence-type tests but not for children from low socio-economic backgrounds, whose learning ability was more effectively revealed on post-tests in a TTT design. As Sewell (1987:440) suggests:

> "The failure of the post-test to improve on the predictive power of the pre-test for middle-class children suggests that, on static measures of ability, middle-class children were

> already performing near their optimal level of learning ability on the pre-test; thus, the impact of the post-test following training was negligible".

This point is spelled out in some detail here as it is fundamental to a central argument in this study. What has been described as 'static' assessment approaches, such as those found in traditional intelligence testing, or achievement tests such as traditional school-leaving examinations, do not fully reveal the learning ability of educationally disadvantaged students. They do, however, perform this task for advantaged students. A context in which both types of students are found, and where there is competition for limited resources in the form of places in higher education, must therefore use both approaches.

A second pivotal figure in dynamic assessment theory building is Feuerstein, who distinguishes between the kind of learning that occurs through trial-and-error, and mediated learning, where experiences (stimuli) are selected, structured and interpreted for the learner by another individual (Hamers & Resing 1993). In this view, mediated learning experiences are not simply important in their own right, they are crucial in enabling a child to learn in other situations. An absence or very low number of mediated learning experiences "... manifests itself in a reduced level of modifiability, a passive attitude towards cognitive tasks, an absence of motivation and a negative self-concept" (ibid:34). Feuerstein developed what he called a 'Learning Potential Assessment Device" (LPAD) to gauge the level of performance he believed individuals could have reached had they experienced the required remediation.

A major difference, within the diagnostic situation, between the work of Feuerstein and Vygotsky is the target of their intervention. For Feuerstein, it is the cognitive structures of the individual, and for Vygotsky and other researchers in this tradition, it is the end product of testing, the ZPD that is the focus of interest. That is, Feuerstein was interested in assessing the modifiability of the individual (the greater the modifiability, the greater the responsiveness to instruction), whereas Vygotsky was interested in measuring the highest level that an individual can achieve with support.

How to operationalise and measure this 'zone' has not been satisfactorily resolved. Siegler (1976), for example, counted the number of prompts needed for an individual to be able to solve a

problem, but this creates another problem, similar to that of defining and using a 'gain' score as described below. That is, the number of prompts is related to the level of the ZAD as much as it is to the ZPD – and a cornerstone of the theory rests on the unreliability of the ZAD as a useful predictor of learning ability.

Dynamic assessment procedures which follow a pre/post test design are subject to the same problems as those faced by all such research designs – that is, the difficulty of explaining changed performance on the second measure, and how this change should be interpreted and used. At the least, post-test performance is affected by:

- initial levels of performance. Candidates who perform at a high level on the initial test might not be able to improve their score on the second test for a number of reasons: the intervention was targeted at weaker students and thus did not meet the needs of more advanced students; it is arguably more difficult (or can take longer) to develop significantly higher levels of skill from a higher base; and statistical regression effects which result in bringing outlying performances (very high or low scores) of an individual closer to the mean of that individual, and hence possibly, to the overall mean.
- the effect of the intervention.
- the inevitable practice effect arising from completing two comparable kinds of tests within a short space of time. If the pre- and post- tests are not comparable, this last confounding variable will be avoided, but a new unknown one of comparability will arise. If the time between tests is lengthened in an attempt to lessen the practice effect, other variables will enter the picture, such as maturation or incidental learning. In addition, of course, lengthening the time between tests will substantially add to the logistical problems and costs associated with these procedures.

Because of these difficulties of interpretation, it is not easy to determine which score should be given most weight: the 'gain' score (i.e. the improvement in scores between the pre- and post-tests), or the post-test score alone. Guthke (1993) and others, like Embretson (1992) and Hamers and Sijtsma (1993), however, conclude on the basis of several studies involving gain scores, that

the post-test score is the more reliable and valid measure to use as an index of an individual's ability to learn. In short, this view asserts that the ZPD, once rendered measurable, is the predictor of choice.

Several studies and projects have used dynamic assessment principles in relation to selection to Higher Education in South Africa. However, in two of the projects, the TTT and the UNIFY projects, the focus is on curriculum development rather than selection, and thus the predictive validity of the selection instruments is difficult to assess.

As was discussed above, the TTT project at the University of Natal pioneered the dynamic approach to selection in South Africa some years ago, in an interesting although costly admissions project. Candidates (for the Faculties of Science and Arts) were brought onto campus prior to the academic year, and after having written various tests, underwent an intensive two-week, residential, teaching/learning immersion experience. They were tested at the end of the fortnight, and selections made primarily on the basis of the development that had occurred, as expressed in the second testing session.

The aim of the project in the Faculty of Arts, however, was to broaden access through the development of distance learning materials that would effectively prepare applicants for future study, and the TTT project in this Faculty area was not validated in terms of its selection instruments. The TTT project in the Faculty of Science (known as the Joint Selection Project for Science and Applied Science - JSPSAS), in contrast, had as an important aim the development of a reliable selection instrument. As part of its validation process, the relationship between the JSPSAS' dynamic selection procedure and the UCT AARP tests was investigated. The findings revealed a .77 correlation between the scores of candidates on the two tests (the AARP tests written at the beginning of the fortnight and the JSPSAS tests written at the end of the TTT experience). As Loubser (1997:11) suggests in an unpublished JSPSAS report, "... performance on one of the tests would be equivalent to performance on the other". This strong correlation suggests that the AARP tests were tapping into the same 'responsiveness to instruction' ability that the TTT fortnight and testing was designed to do. On the basis of this evidence, recommendations

were made to use the AARP tests for selection at Natal, as they required no expensive intervention. In 1998, the JSPSAS dynamic selection approach was discontinued, largely due to funding difficulties, and the selection instruments were not fully validated.

Zaaiman (1998) reports on the contribution of the Raven Progressive Matrices (RPM) to the predictive validity of the UNIFY selection tests. The Raven study was undertaken in order to investigate whether the dynamic testing principles purportedly underpinning the RPM tests would be more appropriate than a single testing session for selecting the students, all of whom were considered to be educationally disadvantaged. The Advanced Progressive Matrices (APM) test was selected as being most appropriate for the UNIFY group, and two equivalent forms of this test were developed, in order to minimise the practice effect which would occur if the students wrote the same test twice. Version 1 of the APM was written at the beginning of the academic year, and Version 2 five months later, after the students had been exposed both to the UNIFY course content during those five months, and to a specially designed intervention just prior to the Version 2 testing session. Unfortunately, as Zaaiman acknowledges, no control group was created, as could have been done by including students who had not been exposed to the intervention. It is thus not possible to assess whether any improvement could have been ascribed to the intervention or not. The correlation (Cronbach's α) between the two versions of the APM was 0.43 (evidence against zero correlation is at better than 0.001 level of significance). As Zaaiman points out, this correlation is lower than would be expected if the same constructs were being measured at each time of by each version, and was possibly due to the time lapse between the two tests which allowed significant other variables to intervene.

In terms of predictive validity, neither of the two APM tests, nor the 'gain score', correlated significantly with the students' final UNIFY results. Nevertheless, interesting (although not statistically significant) information emerged which points to the usefulness of the post-test (version 2) score being greater for students who scored in the lower range on the pre-test (version 1) than the pre-test score itself. It thus supports the use of the post-test score, although this support must

be regarded as highly tentative given the absence of statistical significance of the results. It was decided not to pursue the idea of a dynamic selection test for UNIFY further, on the grounds of:

> "... the non-significant results, the existing uncertainty in the predictive validity of dynamic assessment instruments ..., and the logistical problems involved in implementing a dynamic assessment method for large numbers of students" (Zaaiman 1998:172).

The Alternative Admissions Research Project (AARP) in South Africa, introduced briefly in 4.3.2 above, attempts to incorporate the intervention phase into the admissions tests themselves, on the grounds that costs and logistics prohibit a true test-teach-test design. This research is reported in detail in Chapters Eight - Twelve. In summary, the AARP intervention is provided through the technique of task scaffolding, where each task builds on the previous one, and where students are actively involved in constructing responses to carefully designed and supported tasks. In a study of the approach, Yeld and Haeck (1997) conclude that task scaffolding elicits different performances from comparable candidates to those elicited by more traditional approaches, but does not simply make the test easier for all candidates - in fact, the gap between weaker and stronger AARP candidates actually widens – and should do so if in fact the instrument is to signal prospective benefit if the appropriate candidates are selected.

4.5 Conclusion

It has been argued in this chapter that for educationally disadvantaged students, achievement tests, based as they are on assumed previous learning, are particularly punitive. From the point of view of selection, they are not particularly useful, as results tend to cluster in the lower ranges. Tests designed to tap core competencies will continue to advantage privileged candidates, as opposed to educationally disadvantaged candidates, almost as much as curriculum-aligned tests. That is to say, while an improvement on achievement tests on the grounds of fairness, core competency tests still reflect candidates' educational opportunities. Indeed, no kind of test can ignore the differences between the two groups if the disparities in educational provision are sufficiently great.

Because of this inherent feature, it is argued in this chapter that non curriculum-aligned, core skills tests which are developed as far as possible on dynamic lines represent the most effective approach to assessment in terms of assisting talented yet educationally disadvantaged students to demonstrate their ability. In addition, it is argued that such tests can have a dual function. That is, they can serve as a source of additional information about the abilities of educationally disadvantaged students, and, in addition, they can serve as a quality assurance indicator running alongside national achievement tests. The importance of traditional achievement tests, however, should not be undermined, as they yield essential information about what the system is or is not delivering, and where interventions and improvements are necessary.

Chapter Five focuses on the testing of core skills, paying particular attention to different understandings of knowing and learning.

CHAPTER FIVE

LITERACY, ACADEMIC LITERACY AND ASSUMPTIONS ABOUT KNOWING AND LEARNING: IMPLICATIONS FOR ASSESSMENT

5.1 Introduction

Chapter Four concluded that selection for Higher Education in South Africa should, for the foreseeable future, include two types of tests in addition to other information such as students' educational backgrounds, rural/urban origins, and so forth. These two types of tests are (i) curriculum-aligned (i.e. achievement) tests, and (ii) competency tests in core skill areas. It was argued that the school-leaving examination, while currently flawed, is the logical vehicle for the first of these types of tests, and that improvements in the teaching and learning preceding the examination, as well as much-needed technical and procedural improvements related to the examination itself, would make this examination legitimate, useful, and acceptable.

It was previously argued (see Chapter Three in particular), that an essential requirement of test fairness – particularly in relation to achievement tests - is 'opportunity to learn'. Until the unacceptably high degree of heterogeneity in educational opportunity (i.e. opportunity to learn) has been substantially reduced, therefore, the use of external, curriculum-aligned achievement tests as the sole or even major determinant of future educational opportunity is not legitimate, and would unfairly discriminate against educationally disadvantaged learners.

It is, of course, acknowledged that all tests are, to some extent, curriculum-aligned in that past learning will influence performance through helping to shape the way candidates approach and respond to tests. However, the less closely a test is aligned to a particular curriculum, the less will performance on that test be tied to achievement in that curriculum. Thus, it can be argued that tests based on competencies in core skill areas that are non-subject-specific will reduce the extent of the disadvantage experienced by learners who have had poor prior learning opportunities.

For this reason, the second type of test was proposed as a possible additional indicator of candidate merit. This type of test, however, is controversial in that since it is deliberately not tied to any specific curriculum, there is no existing agreed basis for testing. The content and emphases of the test/s would therefore need to be comprehensively researched. In addition, while requirements of efficiency and affordability point to the need for the number of competency tests to be kept to a minimum, the skill areas covered in the tests must be as comprehensive as possible in order to maximise the information that can be derived from test performance, and to ensure adequate sampling of skills.

An example of an established test which aims to test competencies in core skill areas is the Australian Queensland Core Skills (QCS) Test. This test is the second of the two high-stakes assessment regimes administered in the State of Queensland senior schools, the first being a carefully and comprehensively moderated system of internal, subject-specific, school-based achievement tests, known as the Senior Certificate examination.

The QCS test is used to generate an individual result for a student, as well as to provide important indicators of group performance. In this latter role, it serves to provide essential information to facilitate overall moderation of the school-based assessments. This moderation is achieved in the main on the basis of the group parameters obtained from the QCS (particularly 'average' and 'spread'), by scaling both the scores obtained between different subject groups within individual schools and between schools. The test is compulsory for all students registered for full-time study at Queensland senior schools, but may be taken on a voluntary basis by other candidates.

In brief, the QCS assesses achievement in 49 common curriculum elements, identified as underpinning the SC curriculum. In other words, the test is 'grounded' in the senior curriculum, while not being aligned to any specific subject. As the Queensland Education Department states, "It tests the skills learnt from the combination of subjects in a balanced curriculum" (Queensland Board of Senior Secondary School Studies, 1999).

A sample of the 49 'common curriculum elements' (the cognitive skills that underpin the senior curriculum) is listed below: recalling/remembering; interpreting the meaning of words or other symbols; interpreting the meaning of tables or diagrams or maps or graphs; summarising/ condensing written text; recording/noting data; estimating numerical magnitude; calculating with or without a calculator; structuring/organising extended written text; expounding a viewpoint; analysing; synthesising; and extrapolating;

The importance of core cognitive skills is also highlighted by the South African Qualifications Authority (SAQA), in its identification of the following critical outcomes[26] that should underpin all qualifications registered on the National Qualifications Framework (NQF). In brief, learners should be able to successfully demonstrate their ability to:

- communicate effectively using visual, mathematical and/or language skills in the modes of oral and/or written presentation;
- identify and solve problems by using creative and critical thinking;
- organise and manage themselves and their activities responsibly and effectively;
- work effectively with others in a team, group, organisation and community;
- collect, analyse, organise and critically evaluate information;
- use science and technology effectively and critically, showing responsibility towards the environment and the health of others;
- understand that the world is a set of related systems, and that problem-solving contexts do not exist in isolation; and

[26] Glossed by the Department as "generic, cross-curricular, cross-cultural outcomes".

- show awareness of the importance of effective learning strategies, responsible citizenship, cultural sensitivity, education and career opportunities and entrepreneurial abilities (DoE 1997a:16).

While not all of these outcomes are demonstrable in a formal testing situation, it is clearly envisaged by the curriculum planners that continuous assessment will ensure that the outcomes are all covered during, and as an integral part of, schooling. However, even if the great difficulties posed by continuous assessment in terms of moderation are set aside, it is highly unlikely that these challenging outcomes will soon, or in any convincing way, be reflected in assessment practices in South Africa (Taylor & Vinjevold 1999, Jansen 1997, Pahad 1996). As Chilisa (1999:28) remarks in an assessment of similar developments in the national examination system in Botswana, "Teachers report difficulty constructing test items that assess the higher levels of thinking assumed by the objectives. They also experience difficulty constructing assessments that evaluate students' usage of scientific process skills ...". Such difficulties are also reported by Taylor (1999:203), who concludes that "... the more sophisticated forms of assessment advocated in the new curriculum are well beyond the reach of the majority of South African teachers at this stage". Pahad (1996:18) notes that "... although the policy debate continually touches on assessment, it has failed to acknowledge its central importance to the envisaged transformation of education and training in South Africa".

Any assessment that purports to investigate learning and understanding must be grounded in defensible, up to date assumptions about the nature of knowing and learning. While this assertion holds for all kinds of assessments, it is particularly essential in as ambitious an endeavour as cross-curriculum, cognitive skills testing. This inference is made partly because curriculum-aligned achievement testing, by definition, has already taken many of the decisions about what needs to be tested. In reality, of course, assessment and curriculum are iterative, and developments in assessment technologies can promote – or constrain – developments in curricula, as well as vice versa. However, while testing will always involve content, core skills testing is less directly dependent on it. At the risk of overstating the position, it could be argued that when assessing non-subject-specific core cognitive skills, the very essence of what is being assessed is what it is

to know and be able to do something, and not directly what it is that needs to be known and 'done'. Such testing therefore places a premium on sound assumptions about the nature of knowing and learning.

In addition, a sound understanding of how students learn is essential if one aims to go beyond assessing what students know or have been taught how to do, i.e. to tap 'potential' and not only manifest ability. Chapter Four discussed this issue in some detail. One of the reasons for considering the possibility of introducing an additional test or tests in core skill areas to the repertoire of tests for selection to Higher Education concerns the issue of test fairness. As was discussed in Chapter Three, 'opportunity to learn' is a major factor in the determination of whether a test is fair or not for all candidates. It was concluded in that discussion that the greater the inequities in preparation, the more unfair are tests which are based on a curriculum, as some candidates would have covered the curriculum, and others not. Thus, including a test of non curriculum-specific skills and knowledge would lessen the degree of disadvantage directly attributable to previous instruction. This issue is taken forward in both this chapter and Chapter Six, which discusses the impact of educational disadvantage on learning and knowing in relation, particularly, to the area of academic literacy.

The core skill area that is recommended in this study as appropriate for the development of an additional test is that of academic literacy. Literacy and literacy practices are intimately related to what it is to know and to do in modern society. It is important, however, to stress the controversial and politically sensitive nature of literacy as the term is understood in modern societies, before narrowing the discussion to the sub-field of academic literacy.

5.2 Literacy and Academic Literacy

In the United States, the National Adult Literacy Survey (NALS) of the literacy skills of adults (16 to 64 years of age), was conducted in 1989/90. Panels of experts developed the following definition of literacy: "Using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential". This emphasis on literacy as a tool to use for life's

challenges, and to promote personal development, is necessarily very general. Nevertheless, it accords well with the critical outcomes identified in the SAQA document referred to above.

The NALS definition of literacy was developed to cover the literacy needs of all adults in American society. Academic literacy, however, refers to a more specific set of needs, and is understood in this study to refer to those aspects of literacy required by contexts of learning and teaching which are highly dependent on reading and writing as vehicles for meaning construction, and whose context is customarily that of formal education. The following discussion of literacy is based, in the main, on the area of academic literacy.

Contemporary understandings and uses of the term 'literacy' reflect quite different ideological positions, as well as almost diametrically opposed views, on knowing and learning. One approach to literacy understands literacy as a 'technology' (Goody 1968, Olson 1977, for example) – a set of techniques for decoding and encoding written and auditory text. Its origins and many of its theoretical underpinnings can be found in behaviourism. The second, more recent, view of literacy stresses "... the essentially social character of text ..." (Hill & Parry 1994:21), and claims to be grounded in cognitive and situative perspectives on knowing and learning.

- An illustration of the power of the notion of 'literacy' is offered by the way in which the results of the NALS project were used. NALS developed three proficiency scales to measure what were viewed as three distinct literacy skills: prose, document, and quantitative. These scales, or dimensions, are conceptualised as continua, so that individuals are not judged as 'literate' or 'illiterate', but as demonstrating various levels of literacy skills.

Results from this survey, as had others before it[27], shocked Americans. The intense public interest in the results of this and other surveys reflect the reality that "... literacy has become a primary indicator for judging national progress, for granting opportunities for access and advancement, as well as for the allocation of rewards" (Kirsch, Jungeblat & Campbell 1992:1). As such, it is a powerful instrument in the shaping of public policy, particularly but not only insofar as education is

concerned. One of the consequences of the NALS study was to promote the 'standards-led reform' movement in education, which is closely tied to mandated state-wide standardised testing.

A major contribution of the New Literacy Studies (NLS) movement, the second of the two views on literacy mentioned above, has been the highlighting of the way in which the notion of literacy is manipulated by powerful players in society. It is important to understand the need for a clear understanding of this rhetoric, as different understandings of the problem, if indeed there is one, require very different responses. If there is no genuine crisis, however, attention, and scarce resources, could be directed elsewhere.

Crisis rhetoric about literacy, Welch and Freebody (1993) argue, recurs at different times in educational circles, and has been with us for a considerable time. It is visible, for example, in media headlines along the following lines: "Nine out of ten teachers fail spelling test" (*The Guardian* 7 July 1997), "SA students are scientifically illiterate" (*Mail and Guardian* 18 January 2000). Four possible explanations offered by Welch and Freebody for the emergence of a so-called literacy crisis are discussed below.

The first explanation is that there has been a genuine decline in literacy standards – i.e. that there is indeed a crisis. Numerous studies, however, have indicated that, in 'developed' countries at least, no evidence of a decline in literacy performance has been found (Masters & Forster 1997, Kellaghan & Greaney 1992). The second explanation is that literacy demands have changed or increased and that educational responses to these changes have not been effective. The need for employees to be 'computer literate', for example, has undeniably expanded the scope of what it is to be 'literate'.

Welch and Freebody's third possible explanation for the recurrent emergence of 'literacy crisis rhetoric' is that employment conditions have resulted in changed demands from employers for higher formal educational qualifications, for example, although the literacy demands of jobs might not have changed. The fourth explanation offered by Welch and Freebody for the recurrence of a

[27] For example, *A Nation at Risk*, a report by the National Commission on Excellence in Education (1983), which

so-called crisis in literacy is that the crisis is a 'confection' – a means by which attention can be diverted from other (real or imagined) problems, or by which these problems can be explained.

Street (1997), however, turns to an examination of developments in NLS themselves as a possible cause of current characterisations of literacy and literacy practices as in crisis. These characterisations, he suggests, are a direct response to

> "...recent theories that take a provisional, interpretative, dialogic view of language, literacy and education – a perspective that continually questions authority and exposes the contested and ideological nature of language and literacy practices" (op cit:4).

A main thrust of these 'recent theories' – NLS theories - is to question the one-true-meaning approach to language and literacy: that is, to the approach characterised above as a 'technology'. Instead, individual, contextualised interpretations and methods of meaning and meaning construction are valued and foregrounded. In so doing, the power and control of authority is seen to be challenged. Street suggests that it is "'Hobbesian fears'[28] of this flexible and provisional perspective of language and literacy that are the primary catalysts and maintainers of public debates about literacy "crises"".

The theoretical underpinnings of these contrasting views of literacy and literacy practices differ markedly, as could be expected, and are found in different understandings of knowing and learning. Research on learning and cognition covers a vast field, and has a relatively long history. It has not, however, after decades of development and activity, "... settled into a single theoretical account" (Greeno, Pearson & Schoenfeld 1996:5) of understanding and learning. A useful frame for understanding this research is provided by Greeno et al (ibid) through the identification of four perspectives from which to investigate different aspects of what it means to know and understand. The perspectives are: behaviourist, differential, cognitive, and situative.

---

galvanised the USA into a fundamental restructuring of its education system.

[28] Hobbes, in the 17th Century, argued for strong government in order to stave off chaos and disorder, which would result, he argued, in a collapse of civilisation.

5.3 Behaviourist and Differential Perspectives on Knowing and Learning

The behaviourist perspective is based on a perception of learners as receivers of carefully structured and sequenced stimuli, not as active processors of largely self-selected material. The onus is on the environment to provide an effective learning opportunity, and not so much on the learner, whose responsibility lies primarily in taking advantage of what is on offer. The rote-learning that assumes such prominence in poor educational contexts fits in neatly with this view of learning, for both teacher and learner. As Greeno et al suggest, from a behaviourist perspective to 'know' something is to have acquired "... an organised collection of stimulus-response associations" (1996:6). The application of this knowledge in novel situations is not prioritised. It should be noted, however, that from a cognitive perspective, it is also accepted that certain elementary skills, facts and concepts can be (and ought to be) acquired and automated. For example, knowing the multiplication tables 'off by heart' is of great value in that conscious attention can then be directed to more complex aspects of arithmetic performance. Research in the behavioural perspective, however, assumes that all complex behaviours can be broken into their constituent elements and learned as such[29]. When sufficient prerequisite elements - or skills - have been acquired, learners are deemed ready to move on to more complex levels. The following view of learning is in stark contrast:

> "... understanding is not cued knowledge: performance is never the sum of drills; problems are not exercises; mastery is not achieved by the unthinking application of algorithms. In other words, we cannot be said to understand something unless we can employ our knowledge wisely, fluently, flexibly, and aptly in particular and diverse contexts" (Wiggins 1993:200).

Both behaviourist and differential perspectives are based on what Resnick and Resnick (1992) have termed 'decomposability' and 'decontextualisation' assumptions. As the terms imply, these assumptions rest on a componential notion of knowledge and skill: that is, that knowledge and skill can be broken down into independent pieces and, consequentially, taught, learned and assessed as independent pieces. As Resnick and Resnick (1992:42) suggest, such a view of knowledge

[29] The differential perspective, while similar in many respects to that of behaviourism, is primarily concerned with the 'amounts' of ability and knowledge that have been acquired by learners. The assumption is that "[W]hatever exists, exists in some amount and can be measured" (Greeno et al 1996:6). How this assumption is conceptualised structurally is through the idea of traits, aspects of knowing in specific domains or sub-domains, for example, 'reading comprehension' or 'word attack skills', respectively.

and skill "... supports a notion of teaching and testing separate skills, reserving their composition into a complex performance for some indeterminate later time". As is discussed below, this view has been seriously challenged by recent research, which has confirmed the importance of the links and interactions between skill components.

In addition, the decontextualisation assumption emphasises the generalisability of knowledge and skills, that is, their context independence. In this view, the context in which a skill is developed or acquired is largely irrelevant. Thus, a student who learns what counts as evidence, and how to use this evidence to support arguments, in historical studies would also know what and how to employ evidence in Physical Science, that is, the skill is 'transferable'. As Moore (1997) suggests, discussion about generic, transferable skills commonly assumes the following: that skills can exist independent of context (and can be assessed independent of context); that they are straightforward and mechanical; and that separate skills-based courses are the most effective vehicle for teaching them. However, as Wittrock (1991:11) notes, "transfer does not seem to be a natural or automatic occurrence ...", but on the contrary to require a high degree of metacognitive awareness as well as effective use of learning strategies. In other words, transfer seems to occur most effectively when learners have learned not only the skill, but also to be aware of their thought processes, to be able to manage their learning, and to actively seek ways to relate this new knowledge to other areas of learning, including their prior knowledge and the knowledge used in their everyday lives. This understanding of transfer, and the implications of this understanding for how best to teach for transfer, is supported by an extensive body of research, such as that on the use of comprehension strategies (Brown, Campione & Day 1981), the use of mental models in problem-solving (Mayer 1989), and metacognitive training (Palinscar & Brown, 1984).

A further challenge to the sustainability of the decontextualisation assumption comes from research into expertise which has focused on the critical role of domain-specific knowledge in the execution of complex skills (Allard & Burnett 1985, Chi, Feltovich & Glaser 1981). This challenge is based on a recognition that the acquisition of complex skills typically requires one to master a

substantial amount of detailed knowledge, and that much of this is specific to a particular domain. For example, research by de Groot (1965) into the ability of novice and expert chess players to remember the positions of chess pieces after a very brief opportunity to study the boards, demonstrated convincingly that the chess masters could remember more positions - in fact, over four times as many, on average. However, this finding applied only when the pieces were arranged meaningfully, and not when they were arranged randomly. The conclusion drawn from this conditionality was that random arrangements of chess pieces do not draw on chess-specific expertise in the same way, or to the same extent, as do meaningful arrangements in the domain of skill. Thus the reduction of the difference between novices and masters followed the reduction of the importance of chess-specific expertise.

From the point of view of transferability, this particular piece of research suggests that what was transferred to the memory task was the ability to recognise patterns and thus simplify the task, and that when this ability became irrelevant, the masters lost their advantage. In other words, the closer the task to the domain of expertise, the more likely it is for transfer to take place. It is interesting to note, moreover, that despite the countless hours spent by chess masters in examining the position of pieces on a chequered board, they were no better at simple recall than novices, even though at a common-sense level this simple recall might have been assumed to be part of the skill of playing top quality chess. What this finding points to, it could be argued, is the danger posed by assumptions about transferability, in that the enabling skill is not always straightforward to identify, and indeed might only 'exist' in combination with other skills and in a certain relationship with a specific domain. As Glaser (1991:23) suggests, "Concepts are bound to procedures for their application and to conditions under which these procedures are useful".

In response to Resnick and Resnick's critique of assumptions on decontextualisation, Messick (1994) points out an inescapable consequence of the view that skills are specific to consequences, or procedures, as Glaser suggests. In insisting that knowledge and skill cannot be separated from the context in which they occur, an extreme form of 'situated cognition', he argues, Resnick and Resnick's position leads inevitably to "... a behaviouristic proliferation of skill constructs..." that

make inferences from one instantiation to another impossible to defend. The solution to this, Messick suggests, is to conceive of 'cross-contextual' assessment, which holds that "... what is important is not that the skill appears different in different contexts, but that it changes non-randomly with conditions and hence correlates with construct-relevant variables" (op cit:18).

In order to achieve this, Messick suggests several assessment approaches to minimising the effect of interactions with context. These approaches include attempting as far as possible to downplay the effect of context - for example, by deliberately choosing themes about which candidates know little - or using a range of item types/formats and contexts to tap the same skill, and checking for consistencies in response patterns.

5.3.1 Literacy from Behaviourist and Differential Perspectives

As was outlined above, recent debates about the nature of literacy and literacy practices have focused on two major perspectives. The first of these understands literacy as a technical skill or set of skills (what many theorists have called 'a technology'), consisting mainly of the ability to 'read' and 'write' - i.e. to decode and encode in a largely mechanical fashion. The second understands literacy as embedded in ideology and social practice. This latter understanding of literacy is discussed under point 5.4.1 below, along with its implications for assessment.

Proponents of the so-called autonomous model of literacy would not go so far as to deny that different social contexts impact powerfully on the uses to which literacy can be put. However, by and large, they would not support the notion that reading and writing themselves can be considered to be social processes. An example of such an understanding of literacy can be seen in the approach taken by the NALS project, described in 5.2 above. In order to assist in the interpretation of performance on the three literacy scales (document, prose and quantitative), 'benchmark' tasks were selected along each scale. These were defined in terms of task complexity: for example, the number of pieces of information in the question that needed to be matched to information in the document, or the number of distractors in the document. The benchmarking process, however, did not include the prior knowledge brought to the task by the reader, or in fact any task taker characteristics of the kinds outlined in 5.4 below. It is possible,

therefore, that the NALS project considerably underestimated the literacy levels of some categories of respondents, particularly those for whom literacy is associated almost exclusively with work environments. Had the benchmark tasks reflected contexts with which respondents were familiar or which were experienced by them as legitimate, the performance of those respondents would have resulted in higher demonstrated achieved literacy levels. It needs to be borne in mind, however, that a national survey is, by definition, forced to develop generalisable benchmark tasks. Meeting the needs of one group of workers would inevitably prejudice those of another, and so the test developers have no option but to provide as wide a range of benchmark tasks that are as context neutral as possible. In so doing, of course, they become vulnerable to criticisms of the kind outlined above. In short, these test outcomes may only define minimum, rather than accurate, levels of skill.

This 'autonomous' understanding of literacy rests, as suggested above, on the notion of skills that exist independently of context or purpose. The whole area of what 'skills' are, however, and how they are best developed is itself strongly contested. Current public debates about education, particularly about outcomes-based education, frequently refer to 'life-long learning' as a skill and assume the existence of 'generic', 'transferable' skills. As anyone who has had any experience of teaching a course whose main aim is to equip students with the skills they need in other courses will attest, however, the act of transfer is a goal seldom achieved. An example of the fallacy of this assumption can be found in the vexed area of teaching students how to acknowledge sources. The issue is not the skill of referencing that needs to be taught or learned (i.e. the conventions governing academic attribution), but rather the issues of ownership of knowledge and more importantly, of the way that different pieces and sources of information can be integrated and synthesised. These cannot be effectively taught or learned as a decontextualised skill, which will transfer to other contexts: they cannot be taught as such in any other than the most trivial, surface way. Once again, the role of expertise – in this case of a high level of mastery of domain-specific knowledge – is a prerequisite for competent deployment of the 'skill' of academic attribution.

The 'literacy as technical skill' school of thought tends to ignore or downgrade previous learning and life experiences of learners, and to emphasise progress through a programmed "... sequence of learning modules toward an ultimate goal which can be measured" (Draper 1994:7). Reaching the predetermined goal is all-important, and any questioning of the nature and value of the goal is seldom encouraged. Bourdieu (1984) suggests that the unquestioning acceptance of literacy standards and goals (by those seeking to acquire them) is a consequence of what he calls 'symbolic violence'. This term refers to the ways in which the disadvantaging effects of schooling are inflicted on, and experienced by, those whose cultural capital which is incompatible with success at school. He suggests that this 'violence' is agreed to by the dominated classes when they "... accept the stakes offered by the dominant classes" (ibid:165).

In South Africa, as in many of Britain's former colonies, the position of English in public, commercial and educational life provides an interesting example of symbolic violence. Despite the fact that there are eleven official languages, and despite the weight of research evidence from a variety of contexts which shows the benefits of mother tongue instruction, particularly in the early stages until important cognitive developments have taken place and been consolidated (Plüddemann, Mati & Mahlalela-Thusi 1999, Rodseth 1995, Harley, Allen, Cummins & Swain 1990, Macdonald 1990), there is increasing use of English as the medium of instruction in the early stages. This increase is largely the result of pressure from parents, who have internalised the symbolic equation 'access to English = access to resources'. As Vinjevold (1999:220) concludes on the basis of research conducted by the National Language Project, parents who choose English-medium schools for their children are perceiving English not "... as a language but as a resource. Delaying acquisition of the resource is incomprehensible to parents". Thus, parents and the public continue to demand that this resource be provided, despite the difficulties experienced by learners in acquiring initial literacy in a language that is not their first – difficulties which must contribute to the high drop-out and repeater rates in the early years of schooling reported by

Crouch (1999) from the South African context, and Kellaghan and Greaney (1992) from their study of fourteen African countries[30].

5.3.2 Implications for Assessment

A view of literacy as a technology leads logically to a view of reading and writing as technical skills (Hill & Parry 1994). In turn, this inference has serious consequences for literacy assessment. As suggested above, if it is assumed that learning can, through a series of decompositional analyses, be broken down into ever smaller elements, it is a small further step to assume also that it is appropriate to assess it in that way. Then, having assessed the small elements, assumptions can be made about whether candidates can perform the more complex tasks that would, in this view, result from the sum of all these small parts.

Clearly, the process of decomposition needs to be based on a precise and comprehensive understanding of the knowledges and skills that individuals should be able to demonstrate in whatever domain it is that is being tested. In general, the identification of these behaviours (the decomposed 'elements' referred to above) is carried out by a group of domain specialists (such as curriculum developers or subject specialists) who decide on what behaviours are appropriate for a particular test and/or testing purpose. For example, an academic literacy admissions test for Higher Education would focus on so-called 'knowledge domains' such as 'reading comprehension', and 'writing proficiency'. Within these domains, the specialists would have to decide which sets of behaviours (traits) individuals would need to be able to exhibit in order to be deemed sufficiently knowledgeable. Within 'reading comprehension', for example, the specialist group might include 'understanding relations between parts of text through devices of cohesion', and 'deriving the meanings of words from context'. At this point, test developers would have the task of constructing sets of items that "represent the significant components of the hypothetical domain" (Greeno et al

[30] In this, they are in a similar position to many others in the world. As Lin (1999:393) suggests, from the Chinese context, "[A]ccess to English (or lack of it) often affects the social mobility and life chances of many children and adults who do not speak English as their L1 or L2". The globally dominant position of English, as well as the multi-lingualism of many developing countries, has ensured that acquiring formal literacy and learning English are virtually synonymous. That is, in a context where a multiplicity of languages exists, it seems inevitable that one language will be chosen as a *lingua franca*. The global hegemony of English, recently reinforced by its dominance of the worldwide web, and indeed of the global economy, has made English a common choice of *lingua franca*, particularly since it is rarely the first language of any indigenous group and thus will not exacerbate tensions between these groups. The conflation of 'becoming literate' and 'learning English', however, raises a number of questions about the role of English in reproducing social inequalities in different contexts and parts of the world (Pennycook 1994).

1996:11), the traits, so that performance on these items can be said to represent the degree of mastery of the traits. One of the difficulties of this approach is that while information is obtained about how proficient students are in terms of particular traits (finding meanings from context, or at pronominal reference tasks, for example), it is not clear how much light is shed on how good at reading they are, as the relationship of these 'enabling skills' (Weir 1988, 1990, for example) to the more complex, overarching skill of reading is not yet clearly understood. Candidates' ability to read can only be inferred from their performance on specific items, and that it is not reading ability itself that is being measured. Scores on these traits at present tend to be aggregated, or in some cases, candidates' performances are reported as 'skill profiles'.

The belief that "literacy skills can be isolated from the personal and social characteristics of readers is basic to [traditional] reading tests" (Hill & Parry 1994:19), which are anchored in behaviourist views of learning. This belief results, as was illustrated in the discussion above of the NALS project, in the perception of background or prior knowledge as largely irrelevant to the skill of reading. Interestingly, however, even though such tests are deemed appropriate for candidates irrespective of their knowledge and experiential backgrounds, there is tacit recognition that performance will in fact be affected by variations in background. This can be seen in the deliberate choice of texts of which it is assumed candidates will have little knowledge. This choice is in acknowledgement of the importance of existing knowledge, and the unfamiliar content represents an effort to downplay its significance. Despite this acknowledgement, however, it is assumed that the choice of unfamiliar material will ensure that what is measured is reading skill, and thus such matters as the interactions between the texts in reading tests and test takers' interests, and prior knowledge, are not addressed.

5.4 Cognitive and Situative Perspectives on Knowing and Learning

Cognitive and situative theories of learning, in strong contrast to theories arising out of behaviourism and differentialism, view learning as taking place most effectively through the interactions between the learner and that being learned, and hold that what the learner brings to

the learning situation shapes and impacts in profound ways on the learning experience. Such approaches do not downgrade the importance of input, but highlight the roles of prior knowledge and learning experiences, as well as the roles of other participants and the context of the learning process. In addition, conceptual understanding is considered far more significant than the acquisition of individual facts or pieces of knowledge. Indeed, as Glaser (1991:28) notes, "... the emphasis in theories of learning has shifted from the accumulation of facts and their reinforcement, to the structure and coherence of knowledge and its accessibility in problem solving and reasoning".

Research in the cognitive perspective has given rise to a more 'elaborate and differentiated' view of what it means to know something. This view is organised by Greeno et al (1996) into five "aspects of knowing", which are discussed below with particular reference to literacy. Some of the implications for assessment which relate directly to these aspects are discussed here, although more general discussion of the implications for assessment arising from cognitive and situative perspectives takes place in 5.4.2 below.

The first of these aspects of knowing concerns the roles of elementary skills, facts and concepts. These are seen as basic resources that have been automated, thus allowing attention to be directed to more complex tasks. The example of the multiplication tables given above is relevant here for the domain of arithmetic, as would be knowledge of the alphabet, or syntax and semantics, or of various text conventions in written or auditory text, for literacy. Research has confirmed that "fluent enabling skills, such as reading, spelling and computing, are fundamental to knowledge acquisition" (Goldman 1997:372).

In terms of this 'aspect of knowing', assessments would need to incorporate a wide range of items, format types and conditions. As the discussion in Chapter Three (3.3.2.1) on the robustness of knowledge demonstrated, many candidates are able to subtract a smaller number from a larger when the numbers are arranged vertically, but cannot perform the same operation (using the same numbers) when the problem is expressed in words (Shepard 1997). The implication here is that learners can perform successfully on that task only when the problem is expressed in a familiar

way, but had little transferable knowledge of the skill: that is to say, they had no robust conceptual knowledge of subtraction, but had mastered a specific subtraction technique. As assessment in the cognitive/situative perspective aims above all to elicit information about the status of knowledge acquired by a learner, it is essential that information on the conceptual level is gained, and not only on particular manifestations of techniques. As Greeno et al (1996:13) state, the focus in assessment has shifted from how much knowledge someone has, to "... providing adequate characterisation of just what is the knowledge that someone has".

The second aspect of knowing relates to the role of strategies and schemata. An example of this can be found in the reciprocal process that governs the reading of text. In this process, individuals interpret a text on the basis of their existing knowledge. In doing so, they both add to, or modify, their existing knowledge. In other words, the information in the text impacts on their prior knowledge, and this in turn impacts on their understanding of the text. How readers do this depends on the strategies they have learned and their perceptions of the task in which they are engaged. For example, some readers might choose to do an initial skimming of the text to form an overall idea, or get the gist, before reading it more thoroughly. This initial skimming will shape how the more intensive reading will take place. For instance, they might decide to skip the introductory sections and move on to a specific section which appears to challenge a previous text on that topic. Research has suggested that the greater the range of strategies adopted, the more effectively reading and learning tasks were accomplished (Goldman & Saul 1990).

The role of metacognition forms the third of Greeno et al's (1996) five aspects of knowing that are relevant to cognitive and situative theories. Essentially, metacognition is a monitoring and evaluating function. When a reader pauses to check whether s/he has understood an argument, or when a writer asks the question "Am I making sense here?", they are engaging in a metacognitive activity. Once a problem has been identified, the reader or writer employs a strategy or uses a resource (e.g. consults a dictionary, or a mentor, reads the text again, seeks additional information, or abandons the task). Planning and acting on the consequences of these plans are essential parts of the metacognitive process. Increasingly, research is suggesting that the use of

metacognitive strategies, while essential to effective high-level cognitive functioning, is a learned phenomenon, and highly influential in distinguishing the academic performances of educationally advantaged and disadvantaged students (see Yeld & Hartman 1992, Craig 1989, for example). This point is discussed in more detail in Chapter Six.

Assessing the roles of metacognition and of strategies and schemata poses a serious challenge for assessment. Techniques of elicitation such as think-aloud protocols allow observers to gain some, although still filtered, insights into how learners are constructing meaning. As Greeno et al (1996:32) point out, self-report techniques in this regard are highly suspect, as candidates can learn which strategy they should report, without necessarily being able to employ it, or without fully understanding its significance or relevance to the task. For example, if students are asked "What do you do when you run across a word you don't know?", they are likely to learn the 'politically correct' answers. This difficulty applies also to the assessment of beliefs, discussed below. These processes are arguably most effectively assessed through observations of students working on problem-solving tasks. This type of observation is clearly not possible for all kinds of assessments, however, and it seems that assessment technology has much progress to make in respect of measuring these aspects of learning.

The fourth aspect of knowing relates to the important role of beliefs about learning. For example, if students learn to associate success (e.g. understanding a difficult task, praise from a teacher, good performance on a test, peer esteem) with effort, they are likely to believe that effort is worthwhile, and thus that their actions can make a difference to their learning. Beliefs also impact on learners' assumptions about knowledge and 'truth'. For example, they might predispose a learner to characterise a problem as what Strohm-Kitchener (1983) has called a puzzle (a well-structured problem with only one correct solution) rather than seeing it as 'ill-structured', which could have more than one appropriate response or solution and might require a variety of strategies or combinations of strategies in its resolution.

To use an example drawn from the area of literacy, many readers believe that meaning resides in a text, and that their task as readers is simply to find it. Of course, this notion is not entirely foolish.

After all, one is more likely to come to an understanding of a text by reading it, than by not doing so. It is, moreover, a belief strongly reinforced by most reading tests. However, the belief of many readers that text is a given, not to be questioned or challenged, acts powerfully against their developing into critical, self-reflective readers (Moll & Slonimsky, 1989). Indeed, it seems from research that in order to become effectively literate, several 'habits of mind' are essential. Individuals need, for example, to seek actively to make connections of various kinds, within and across texts and activities. They need to learn to take multiple perspectives (e.g. their own, the author's, other role players), and, perhaps most essentially, to persist until they are fully satisfied that they have understood or achieved their purpose. In a reading test, it could of course be argued that the purpose of the reader is to understand the text sufficiently to answer the questions. The point being made here, however, is that by developing tests which require readers to integrate information drawn from more than one source, and/or to apply information in new and novel ways, the tendency to view 'meaning' as a given in a text would be minimised.

The last of the five aspects of knowing concerns the role of contextual factors. This refers in the main to the knowledge required to interpret contextual features and purposes, and to relate the knowledge acquired in one context to another. Thus knowing, in terms of contexts of knowledge use, refers not only to transfer, but also to appropriacy. This concern with appropriacy is echoed in the field of linguistics, as Alderson et al (1995) note, with the move to redefine linguistic competence (preoccupied with the rules of grammar) as communicative competence (which emphasises the uses that are made of language), and the broadening of models of language proficiency to include not only grammatical, but also textual, illocutionary and sociolinguistic competences.

It can be seen that these 'aspects of knowing' represent a far greater challenge for instruction and assessment than that posed by the model of learning espoused by behaviourism. These challenges are spelled out in more detail section in section 5.4.2 below.

The increasing importance ascribed to context and to interpersonal and social interaction is described by Gee (2000) as occurring across a wide range of disciplines. This 'social turn', as Gee

terms it, started with a preoccupation with individual behaviour in the first half of the twentieth century. It then moved through a focus on individual minds (seen in the cognitivism of the middle to late twentieth century), and is now seen in the new importance accorded to social and cultural interaction[31].

While the theories are closely related, there are nevertheless important distinctions between theories of knowing and learning associated with work in cognitive psychology (e.g. the developmental work of Piaget, or of information-theorists), and those evident in work on situated cognition (Leont'ev 1978, Lave 1996, and others). The distinctions are particularly relevant for this study in terms of their implications for assessment.

Essentially, the distinctions lie in the different foci of the two approaches. The former concentrates on the impact that social and cultural ways of knowing and understanding – as mediated through the technologies, symbols, and forms of language that are appropriated and used by a particular society or culture - have on the thought processes and conceptual development in an individual's mind. In this approach, thinking is a social activity, as it is fundamentally influenced by the society in which the individual finds her/himself. The actual result of the activity itself, however, is essentially individual, and resides in the head of an individual. In contrast, work within the situated cognition approach holds that

> "... knowledge and intelligence reside not solely in heads, but, rather, are distributed across the social practices (including language practices) and the various tools, technologies and semiotic systems that a given 'community of practice' uses in order to carry out its characteristic activities ..." (Gee 2000:181).

In this approach, knowledge and expertise are shared amongst a group of participants, and the individual's level of expertise determines how centrally s/he can participate in the enterprise. 'Knowing' is defined as "... sustained participation in practices involving collaboration and use of resources ..." (Greeno et al 1996:7). While researchers in the situated cognition tradition recognise that all learning adds to and/or modifies prior knowledge within the head of an individual, they hold that much significant learning resides in the interactions between participants in the kinds

[31] Most recently, however, the importance of biological bases of cognition (e.g. Hamer and Copeland 1998) is

of practices defined by Greeno et al (1996) above. This interaction is explicitly identified as a critical outcome by the South African Qualifications Authority (see section 5.1 above) which states that learners should be able to demonstrate their ability to "... work effectively with others in a team, group, organisation and community", and to "... understand that the world is a set of related systems, and that problem-solving contexts do not exist in isolation". Breier, Taetsane, and Sait (1996) offer an interesting discussion of literacy from a situated cognition perspective. Their study describes the tendency of illiterate taxi drivers in South Africa to enlist the help of passengers to decipher traffic and direction signs: a form of 'collective map-reading skills" (J. Muller, personal communication, December 2000)[32].

The difficulties for assessment inherent in this approach can be readily imagined, as tests almost always represent some kind of individualised abstraction from reality. Thus, unless tests are conducted in situations that permit inferences to be drawn about how students perform as interactive participants (for example, by observing students conducting a science experiment and rating them not only on the success of the experiment, but on the effectiveness and appropriacy of their participation in the process), they will not serve as tests which meet the requirements of situated cognition. As Greeno et al (1996:15) state:

> "While the cognitive perspective requires stretching testing technology to capture students' integrative and generative understanding, the situative perspective causes standard on-demand[33] testing technologies to collapse".

This point is taken further in 5.4.2 below. It is worth noting at this juncture, however, that, by and large, the situativists are concerned with informal learning, and with assessing the effectiveness of this learning in a highly context-specific way as is evident in the taxi-driver example cited above, rather than formal learning.

challenging the dominance of social and cultural explanations for how people learn and understand.

[32] The limitations of this strategy are revealed by the authors, who cite one taxi driver who was forced to stop his vehicle on a long trip, as his literate passengers had all fallen asleep.

[33] 'On demand' testing refers to tests which can be written at a test user's or taker's convenience, within reason. For example, the SAT1 test is written on various advertised dates during the year, and test takers can choose which date suits them best. On-demand tests are generally standardised and non user-specific.

5.4.1 Literacy from cognitive and situative perspectives

As was suggested above, the NALS survey aimed to gauge and describe the levels of literacy of all Americans, and thus did not focus on the literacy needs most related to formal education. The NALS definition of literacy is: "Using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential" (Barton 1994:3). The modification contained in the italicised phrases below would expand the NALS definition to include academic literacy. It is not suggested that this expanded definition would be found useful or practical for a project such as the NALS, where the needs of Higher Education, or more generally of formal education, are not a high priority, but that the redefinition is necessary for the purposes of formal education.

> "Using printed and written information to function *effectively at the tertiary education level as well as in society more generally*, to achieve one's goals, and to develop one's knowledge and potential".

The more detailed definitions of academic literacy below come from two universities. The first was adopted by the Academic Board of the University of Sydney, Australia, and appears in its policy document on "Written and oral communication skills of students".

> "Literacy is the ability to read and use written information and to write appropriately in a range of contexts. It is used to develop knowledge and understanding, to achieve personal growth and to function effectively in our society ... Literacy involves the integration of speaking, listening and critical thinking with reading and writing" (Bonanno and Jones 1997:2).

This definition bears a strong resemblance to the NALS version, differing mainly in its specific mention of cognitive development, and knowledge construction. This latter is achieved by omitting "one's" from the phrase – thus "to develop one's knowledge" in the NALS definition becomes "to develop knowledge ..." in the University of Sydney version, a move in line with the core function of a university. In doing so, it tacitly lays a greater claim to ownership of knowledge than does the NALS version, which clearly implies a body of knowledge and skill that exists to be mastered and deployed, and does not explicitly raise the possibility of an individual contributing to that knowledge.

The definition below, developed by the Psychology Department of a South African university, spells out the area in more detail and with greater specificity.

> "Academic literacy is the set of competencies required to think critically, ask questions, communicate and access relevant resources [within the discipline of Psychology] at the tertiary education level. Among these competencies are: the abilities to read complex texts, to communicate through writing, to attend and participate in lectures, to access and use resources including the library, computers and staff and peers, and to write examinations" (Amos 1999:184).

It can be seen that the emphasis here, as in the University of Sydney version, is on literacy as a tool for functioning in an academic context so that effective learning can take place. It is, however, a more focused conceptualisation of academic literacy. It stresses the identification of individuals as agents who must think, ask, communicate, access resources (including computers and people), read, write, attend and participate in lectures. Interestingly, it also includes the writing of examinations as an academic literacy practice. In the emphasis placed on the active participation of the learner in the learning process, this version is closely aligned to the recent findings and theories (discussed above) on how learning and knowing take place, can most effectively be supported, and can be most appropriately assessed.

Both of these accounts accord with the view of literacy as articulated by NLS theorists, viz. that literacy is a set of practices located (situated) in specific contexts. As Gee (1990:46) puts it, "... the study of literacy ultimately requires us to study the social groups and institutions within which one is socialised to interpret certain types of words and certain sorts of worlds in certain ways". This somewhat deterministic view has been challenged by many working in the field of literacy at various levels and in various contexts (e.g. Thesen 1997, Widdowson 1998a, 1998b, Price 1999, Lin 1999). Indeed, more recent NLS thinking has conceded that "... meaning and context are mutually constitutive of each other" (Gee 2000:190), and has acknowledged the role of the individual as agent – a role which was downplayed in the reaction of NLS theorists and practitioners to the preoccupation of cognitivists and behaviourists with the individual.

Nevertheless, the notion of Discourses first propounded by Gee (1990) is useful in understanding how the 'configurations' that result from patterns of behaviour that are typical and often unique to a

culture constrain as well as enable our actions, or ways of being. These configurations include ways of thinking, using space, relating to strangers and intimate friends, signifying pleasure and displeasure, and so forth (again, it is important to note that these are all examples of informal contexts). In relation to academic literacy, Tusting, Ivanic and Wilson (2000:214) suggest that

> "... working in isolation at a desk, using a computer and unlined paper, producing texts with high lexical density, few social actors, little or no reference to emotions and senses, and producing uniform lines of typed black text without graphics are a configuration of literacy practices associated with a traditional academic Discourse".

This 'traditional academic Discourse' has been hegemonic for some decades now, but is under increasing pressure. Indeed, although the description (some would say caricature) above is still valid, an increasingly wide range of literacy practices are becoming acceptable. In NLS thinking, the 'enactive work' (Gee 2000) entailed in creating and sustaining a configuration requires ongoing effort, and is in frequent (perhaps inevitable) conflict with 'recognition work', that is, with the efforts of others to modify or reject the original configuration, or Discourse. At this stage, however, the efforts of NLS theorists have been directed in the main at informal sites of learning. As a result, they have made only a minor contribution to thinking about the ways in which academic literacy, which focuses on the development and uses of literacy in formal contexts of learning, can be assessed. This point is developed further below.

5.4.2 Implications for assessment

As the discussion above suggests, the work of the NLS theorists has great explanatory power for language and general pedagogy. The connection to testing, however is complex. The requirements of reliability, now seen as an integral and fundamental part of validity (Messick 1989) are such that an ordered, stable notion of literacy is essential if it is to be assessed. Too much flexibility and interpretation is simply not feasible, even if it was desirable. As Freebody (1997) argues, principled descriptive accounts of actual literacy practices and demands are urgently needed to replace both current narrow prescriptions and, it must be added, vague, romanticised notions. Clearly, from the perspective of admissions testing, an understanding of Gee's 'social groups and institutions' would assist in the interpretation of certain test performances and in the choice of themes and texts. Equally clearly, however, the most important Discourse in this

situation is that of the receiving institution, and it is the extent to which the candidate can produce a performance interpretable from within that configuration that will count. Indeed, the validation of the tests rests on the relationship between admissions test performance and future academic performance[34].

Rather than simply setting out to measure what students know and can do, cognitive perspectives prioritise assessments and assessment procedures that "... measure how students learn, think, make decisions, and acquire, remember, and apply knowledge gained from instruction" (Wittrock 1991:6). Keywords are 'connections' and 'networks' and not 'quantities' and 'ordered sequences' as prized by the behaviourists.

A major implication for testing arising from the work of cognitive text processing and other cognitive and sociolinguistic research is as follows. Unless candidates are put into situations that require them to find and use (synthesise) information from multiple sources[35], tests will do no more than yield information about 'inert' knowledge (Whitehead 1929, cited in Goldman 1997). The problem is that while fluent reading, writing and computing skills are considered fundamental to effective learning, inferring reading ability by testing understanding of only one text or source says little about how a learner can use her/his reading skills to access other sources of information, and how s/he can integrate the new with the old, or apply the new information in novel situations.

It is clear that a serious assessment challenge is posed by the understandings of knowing and learned held by cognitivists, particularly for large-scale assessments. As Greeno et al (1996:13) suggest, "[T]he conceptual understandings and strategies for reasoning and solving problems that are emphasised in the cognitive perspective are difficult to capture in test items that can be scored simply as being correct or incorrect". Indeed, the relative ease with which learning and knowing from a behavioural perspective can be assessed is largely to blame for the enduring dominance of assessment approaches and technologies that are deeply rooted in behavioural thinking. This

[34] It should be added at this point, though, that as flexible entry routes and curricula are developed, so will the range of Discourses that will be considered to be important widen.
[35] Such information is often presented visually as well as verbally, so that it can form part of the solution to a problem, for example so that it can be made into a coherent argument, or can shape a procedure.

dominance is particularly evident, as suggested above, in large-scale testing operations exemplilfied in the SAT, TOEFL and NAEP tests in the United States, that still rely overwhelmingly on multiple-choice formats which encourage targeting of small, discrete, decontextualised bits of knowledge. For example, the National Academy of Education's (1997) report on the NAEP project recommends that "... particular attention be given to such aspects of student cognition as problem representation, the use of strategies and self-regulatory skills, and the formulation of explanations and interpretations" (Pellegrino, Jones & Mitchell 1999:124) – aspects that they believed past NAEP tests had failed to address adequately. This criticism is not to imply that such tests do not deliver useful and valid information, but rather that certain important aspects of learning are not easily elicited through total reliance on these formats.

Situative theories of learning, too, pose enormous challenges for assessment. It seems clear that the need for advanced and effective collaborative skills (both in distance and contact situations) will become increasingly important as electronic learning and work communities proliferate. Identifying, locating, sharing, analysing and organising information will assume ever more prominence in defining the criteria for effective participation (Hull, Jury, Ziv & Schultz 1994). The mismatch between these abilities and skills and those that can be elicited by current pencil-and-paper tests written by individuals is striking, although again it is important to note that the addition of such predominantly social skills to the more conventional cognitive skills requires the development of testing technologies which complement, rather than challenge, traditional assessment approaches.

In essence, assessment in the cognitive sense involves the elicitation of information about a learner's understanding in a domain in terms of the following two main characteristics (Greeno et al 1996):

- the extent to which it is integrated. Here the focus is on how extensively and effectively the relationships between various items of information or skills and specific situations and problems function. In the subtraction example cited in Chapter Three, Section 3.3.2.1 above, it is clear that many learners could not make a connection between the verbal and arithmetic

forms of the problem; the conceptual schemata for subtraction had not been adequately developed. Forms of assessment that would tap this aspect of knowing would include think-aloud-protocols (with all their attendant difficulties for validity and reliability), flow-charts, concept maps, questions that range across texts, and so on.

- its generative capacity. This characteristic refers to the transferability, that is, the ability to identify relevant aspects of one situation or problem and apply them to another, and to use knowledge learned in one situation in another. This frequently requires some interpretation of the new problem situation, or a reinterpretation or reorganisation of the 'old' situation. For assessment, the challenge is to provide contexts in which students are required to apply, infer and extrapolate from existing knowledge to novel situations. Forms of assessment that would be appropriate in terms of this aspect of knowing would include problem-solving tasks, scenario-building tasks, and open-ended tasks.

In turn, assessment in the situated cognitive sense lays stress on eliciting information about a learner's understanding in terms of the following main characteristics. First, information needs to be elicited about the extent to which candidates are able to work effectively in collaboration with others. Forms of assessment that would be appropriate in this regard include carefully structured group work where the rating system is designed to focus on individual contributions and not simply aggregated group effort. The latter system has given group work a poor reputation in the past, with many educators believing that weak candidates are carried by their stronger peers. This flaw is not inherent to group work, however, but is the consequence of poor task design which is almost entirely product dependent. Shifting the emphasis to the process of group work clearly demands more 'on-task' assessment time, and this demand is a major drawback in times of increasing resource constraints. Another method for gathering information on how effective individuals are as collaborators in a learning task could involve the use of sequenced task components, where individuals have a clearly defined component of a task to complete as an individual but whose completion depends crucially on negotiation with the individuals responsible for the previous and next components. Again, this method would only be effective if clear records (ideally but not necessarily including direct observation) were kept on the interactive, collaborative encounters

themselves, as well as on the impact these encounters have on the component for which the individual is responsible.

It is clear that large-scale, on-demand testing is not easily able to accommodate such assessment approaches. It is nevertheless important to note that the ability to collaborate effectively with others is an increasingly important requirement of learning situations, and that assessments which do not elicit such information are therefore limited.

Second, the extent to which candidates are able to draw on multiple sources of information to promote and support learning is an important aspect of understanding that needs to be elicited. These sources could include peers, instructors, books, use of electronic information resources, including, where appropriate, self- and other-generated web searches, media resources, and suchlike. Learners would be required to demonstrate when and how to use these multiple sources, including when to abandon some approaches and/or sources and generate others. Forms of assessment that are appropriate to elicit this kind of information about learners could most feasibly be provided through continuous assessment, where learners can seek out and use appropriate sources of information, for example by using teacher-provided information packages, the local and/or school library (ideally with internet connectivity), or form study groups. For large-scale assessment the challenge is much greater, particularly in contexts where electricity is not automatically available, and computers are a rarity. Nevertheless, even in large-scale paper and pencil tests, it is usually possible to provide a variety of printed and visual material. The challenge then is to structure tasks which require use of these multiple sources.

A striking feature of assessment within both the cognitive and situative perspectives is the emphasis on diagnosis. The major focus of such tests, according to Wittrock (1991:4), is not to "... provide information useful for comparing achievements of different students, schools, systems, states or nations", but to yield information about the state of learning inside people's heads, so that further instruction and learning opportunities can be designed. However, diagnosis need not be viewed as incompatible with acquiring information about developed abilities.

5.5 Conclusion

Chapter Five has discussed, in the context mainly of literacy and academic literacy, four major theoretical approaches to knowing and learning, and has explicated several of the challenges that these approaches pose for effective and feasible assessments. In the main, it seems unavoidable that large-scale assessment initiatives, in particular, will draw on an eclectic mixture of all the approaches. However, considerations of cost, the need for some form of standardisation, and the difficulty inherent in developing psychometrically robust ways of testing higher-order cognitive skills in particular, tend to favour the development of tests more closely aligned to behaviourist and differential approaches to knowing and learning. Ensuring that cognitive, and where appropriate and feasible, situative, approaches are included requires effort and vigilance.

Chapter Six focuses on one particular arena or type of assessment, namely language testing, with particular emphasis on the testing of academic literacy. As is discussed in the chapter, Language Testing is a particularly challenging undertaking as language is both the vehicle and the medium. The discussion is undertaken primarily from the perspective of cognitive and/or situative approaches to knowing and learning.

CHAPTER SIX

LANGUAGE TESTING AND THE ASSESSMENT OF ACADEMIC LITERACY

6.1 Introduction

Chapter Five argued that the most appropriate area in which to develop additional test/s in the immediate future is that of academic literacy, at a level suitable for the end of schooling/beginning of tertiary education[36]. The discussion of academic literacy in Chapter Five was necessarily brief, and did not directly explore the role of language in knowing and learning. This role is, however, clearly critically important in a context where, for over 80% of the population, the language of learning at almost all levels is not the learners' first language. In addition, the chapter did not pursue the implications, for assessment, of the combination of learning through a language that is not one's first, with the conditions that nurture educational disadvantage. In many contexts, such as that of South Africa, this applies to the majority of learners.

Chapter Six provides an overview of these areas, with particular emphasis on the role of language and language testing in, respectively, developing and assessing academic literacy. The overview includes a discussion of the history of language testing, which reflects changing conceptions of what it is to know and use language. This overview is followed by a discussion of the assessment of academic literacy in a context of widespread educational disadvantage, which has important

[36] It is recognised that a second important core area, namely numeracy, has a similar claim for urgent attention and would complement academic literacy. However, that as the majority of school-leavers do not study mathematics as a

consequences for the development of fair and useful assessments and assessment procedures. In addition, Chapter Six lays the groundwork for the analysis, in Chapters Eight to Twelve, of a currently widely used test in the area of academic literacy. It achieves this through explicating an appropriate construct for an academic literacy test.

Some of the possible causes of educational disadvantage were outlined in Chapter Two. These include low socio-economic status, instruction through the medium of a language that is not the learner's first language, poor school quality, lack of role models, geographic origin, and gender. Clearly, a test of academic literacy has to remain cognizant of all these influences on learning and understanding, and on their implications for the demonstration of learning and understanding. As Carroll (1961:31) states, "[T]he purpose of testing is always to render information to aid in making intelligent decisions about possible courses of action". This emphasis on 'intelligent decisions' goes to the heart of the difficulty faced by tests of language, as "... what is being measured is that most flexible, multidimensional, fugitive, and complex of human abilities, the ability to use language" (Spolsky 1995:39).

Adding to this problem is the difficulty that, as Messick (1994) suggests, language is frequently both the vehicle and the target of the assessment. This further source of difficulty is highlighted by McNamara (1996), who makes a clear conceptual distinction between two kinds of language test. His discussion is located in the context of second language testing, but the strong/weak distinction he draws is equally valid in first language testing contexts. Indeed, in the strong form, the first/second language distinction is hardly relevant, as is argued below.

The distinction between strong and weak forms of language test is based on the nature of the inferences that can be drawn from performances on the test. One kind of test – the strong form - will yield information on an individual's ability to function effectively in future tasks or situations that require the use of language. The other will yield information about an individual's use of language in future tasks or situations, but will not attempt to extrapolate from this information the likelihood of

subject, it is believed that it would be more effective at this stage to include and tap basic numeracy skills as part of academic literacy.

an individual's success or failure in such future situations or tasks. Its focus is more clearly on language as a system rather than the uses to which language is put.

So, for example, the SAT would tend towards the strong end of the continuum proposed by McNamara. It aims to elicit information about candidates' ability to function effectively in English-medium higher education, based on their performance on tasks in the test, irrespective of whether they are first, second, or foreign language speakers of English. Test scores are used by institutions to assist in general admissions decisions. The test is not designed to yield information about traditional language skill areas such as reading, writing, speaking or listening, but rather aim to assess candidates in terms of their ability to apply "... multiple skills and abilities in carrying out a task or tasks" (Hudson 1996:10). Clearly, language is involved in the application of these skills and abilities, but is thoroughly intertwined with other abilities that are considered to be crucial to successful performance in the criterion situation.

Consider the example of a Humanities student writing an essay at the end of a module of instruction, or in a test as the culminating task after a number of related inputs on a topic, such as readings, a tape-recorded lecture, and visual materials. In order to complete this task, the student has to deploy a considerable number of often overlapping skills. For example, the student first has to understand the stimulus materials (the course of instruction or the materials in the test). This understanding will depend on a number of factors, including prior knowledge, interest and motivation, the effectiveness and clarity of the materials/instruction themselves, language proficiency, and the time available to prepare for the task. After this, the student has to understand the task itself, which entails decoding the essay title, or task rubric, as well deploying background knowledge of the formal, rhetorical organisational structures characteristic of different types of texts (Carrell 1988) as a guide in the planning process. After analysing the task, the student has to select certain pieces of information, and reorganise and synthesise these, often with her/his own ideas on the topic. After analysis, the essay needs to be written, edited and revised, and submitted. The difficulty of confidently ascribing strong or weak performance to language proficiency – on its own or even as a major contributor - is evident.

The challenge to language testing inherent in strong language tests is discussed in more detail below, in relation to recent models of language proficiency (e.g. Bachman & Palmer 1996, Cummins 1984), which acknowledge and highlight the importance of cognitive and affective variables in language performance.

The TOEFL, in contrast to 'strong' tests such as the SAT, targets second or foreign language speakers of English, and aims to deliver information to institutions in terms of candidates' English language proficiency. This information is used to admit candidates unconditionally, place them on appropriate language courses, or refuse them admission on the grounds that their low level of English proficiency will prevent them from succeeding in their studies. As Hamp-Lyons and Kroll (1997:1) suggest, it attempts to meet the "... need to ensure a minimum proficiency in English among entering students ...". The test has not been validated as a predictor of academic performance, however. Indeed, it is generally acknowledged that the test reveals little about how students actually use English in academic settings, and admissions officers have reported that many students admitted with high TOEFL scores demonstrate inadequate writing and oral communication skills for effective academic participation (Jamieson, Jones, Kirsch, Mosenthal & Taylor 2000). This outcome is hardly surprising in view of TOEFL's current total reliance on multiple-choice items which, in addition to requiring candidates to select amongst alternatives rather than actually to produce language, aim to target only one element of a skill at a time (e.g. grammar or vocabulary)[37].

In terms of McNamara's 'strong' and 'weak' distinction, the TOEFL is at the 'weak' end of the continuum, in that language is both the target and the vehicle. However, it is used as though it were both 'strong' and 'weak'. In the strong sense, it is used by institutions to admit or reject applicants on the basis that it is able to predict the level of difficulty a student will experience in an academic environment, and therefore the level of academic risk an institution faces if it admits that

[37] This issue here is not the use of multiple-choice, but the emphasis on discrete skills testing – something the TOEFL® 2000 project aims to address. This project is responsible for a comprehensive revamp of the test to bring it into line with customer needs as well as theoretical advances in the fields of second language learning and assessment, and it is believed that these developments will improve the predictive validity of the TOEFL.

applicant. In the weak sense, it is used to inform institutions about the kinds of language-related assistance that their students will require in order to adequately access the academic environment. One of the problems with using a test such as the TOEFL, which has not been developed to reflect the kinds of tasks with which candidates will be confronted in higher education, as though it had been developed for that purpose – that is, as though it were a 'strong' language test - is that it correlates only weakly with academic achievement (see for example Hill, Storch & Lynch 1999, Davies 1988, Criper & Davies 1988, McNamara 1996). As Davies (1988:34) reports in relation to tests of this kind, the "... typical predictive correlation with academic examination criteria is about 0.3".

It can be seen in the examples above that the strong/weak distinction has important implications for test development, and particularly for tests of academic literacy, although it should be recognised that no test is likely to be entirely strong or entirely weak. As McNamara suggests, tests will tend to be 'relatively strong' or 'relatively weak'. This distinction is returned to below.

6.2 Implications (for Assessment) of Changing Conceptions of Language Learning and Use

The role of language within academic literacy is powerful and central, and features prominently in popular conceptions of the reasons for the learning difficulties underlying poor academic performance. While this role is seldom made explicit, the widespread use of tests which are based on language performance of some kind to assist in selection for Higher Education testifies to the common belief that performance in the language that is the medium of instruction is strongly and directly related to an individual's chances of success. The example above of TOEFL illustrates how this belief can lead to an inappropriate use of tests.

In the development of tests of academic literacy, it is essential to bring together two fields: learning and language learning. This combination is particularly important in contexts where the provision of educational opportunities is, by and large, the most inadequate for students for whom the language of learning is not their first language. Chapter Five discussed new ways in which knowing and understanding are conceptualised, and included new theories of literacy in this

discussion. The discussion below uses as a framework the history of language testing to explore notions of what it is to know and understand a language.

The history of language testing, up to and including the advent of the communicative competence movement, was characterised by Spolsky (1975) as having three stages[38]. He named the earliest stage 'pre-scientific', and referred to a time when candidates wrote a general essay, completed an open-ended task, or were interviewed, and the assessment was conducted, often intuitively, by an authoritative, officially designated examiner. Morrow (1981), in dubbing this the 'Garden of Eden' stage, captured its essential innocence and simplicity. Such important concerns as accountability, transparency, reliability, and so forth, were virtually unknown or, at the least, largely ignored. In the primacy it accorded to task (can an individual write a satisfactory essay, respond appropriately in an interview), this stage in language testing bears a resemblance to the performance testing paradigm which is dominant today. However, the relationship of the task to the purpose of the test, the naïve ways in which the task was evaluated by the assessor, and the lack of explicit interest in the nature of the abilities underlying performance, are some areas of profound difference.

The second phase was labeled 'psychometric-structuralist' by Spolsky, and the 'Vale of Tears' by Morrow. In this approach, language learning was envisaged as a generally uniform set of stages, relating primarily to grammatical mastery, and the goal was unequivocally defined by native speaker performance. Indeed, in the development of his Test of Aural Comprehension in English as a Foreign Language, Lado (1946) – a key figure in this phase - focused on items which native speakers would get correct. If they did not get these correct, Lado assumed that something other than language was being measured. On the basis of contrastive analysis, the test focused on structural elements of English, and attempted to minimise the influence of situation, which was regarded, essentially, as confusing the issue. Research on native speaker variability has a long history, however. For example, in 1981 Alderson was asking "... whose performance, which performance is criterial?" (Alderson 1981b:49). His concern followed his study of the performance

[38] He later repudiated this suggestion in favour of a view of the development of language testing as shaped by "... an unresolved (and fundamentally unresolvable) tension between competing sets of forces" (Spolsky 1995:354). These forces, or factors – namely feasibility, usability, and reliability – act, he argues, as "... constraints on the possibility of

of native and non-native speakers of English on a cloze test (a test item format where every nth word is deleted, and has to be supplied by the candidate). While the native speakers out-performed the non-native speakers, the difference was slight, and there was considerable overlap (Alderson 1980). Further research has been conducted by, *inter alia*, Clapham (1996), Hamilton, Lopes, McNamara and Sheridan (1993), Bachman (1990), and Oller and Conrad (1971). These studies point to the influence of educational background and non-linguistic characteristics, particularly but not only in connection with tests that can be characterised as 'strong' language tests in terms of the distinction drawn above, that is, tests which contain tasks that simulate those in the criterion situation rather than focus on language as a system. The studies clearly reveal the flaws in Lado's assumption of native speaker homogeneity, and raise questions about the testing approach that was based on this assumption.

The task of assessment during this stage was to find out what errors were being made by learners along their orderly path to native-speaker-like performance. While the notion of native speaker performance as a legitimate or even wholly desirable goal has been seriously challenged, it remains today, of course, an important aim of assessment to identify learners' problems. The structuralist approach, however, targeted small bits of language (e.g. pronunciation, progressive tenses, contractions), and from there made assumptions about a person's proficiency in the broader domain which the test items purportedly sampled. This breaking down of complex behaviours, or higher order skills, into simple (and hypothesised) components rests on a basic assumption of behaviourism - that a candidate's number of correct responses to items sampling deconstructed sub-skills will reveal how much of the broader skill s/he has achieved. Tests developed along these lines run counter to the findings of cognitive research, which, as Resnick and Resnick (1992:42) suggest, indicate that "...complicated skills and competencies owe their complexity not just to the number of components they engage but also to interactions among the components and heuristics for calling upon them".

developing a valid test" (op cit: 356), and are differently powerful at different times. Nevertheless, Spolsky's 1975 three-stage framework provides a useful heuristic for framing a brief discussion of the early development of language testing.

The growing influence of structuralist views of language was accompanied by the rise of an assessment technology that both supported and promoted it, and has contributed to its tenacity. Elements of this approach can be seen in many large-scale tests still in use today, such as the TOEFL.

One of the side effects of the ascendancy of psychometrics was, as Alderson (1981a) suggests, the rise of language testing specialists, who operated in "... an arcane world of numbers and formulae" (op cit:5). In turn, the dominance of language testing by psychometricians both intimidated and alienated (Lazaraton, Riggenbach & Ediger 1987) many of those involved in language teaching, the majority of whom are humanities graduates, and the field of testing became increasingly distant from the concerns of those involved in language classrooms. What was considered important to measure became that which it was possible to measure technologically and interpret statistically. The rise of multiple-choice format questions and an atomistic (that is, a focus on the testing of very small, discrete elements) approach to test design went hand in hand with a theory of language (e.g. Lado 1961) which assumed that "... knowledge of the elements of a language is equivalent to knowledge of the language" (Morrow 1981:11).

Robinson (1973, cited in Morrow 1981) identified several major consequences of reliance on testing procedures which could be objectively scored, and on the testing of items which had been broken down into small elements. Most importantly, the candidate, in writing the test, actually produces almost no language – in a purely multiple-choice format test, s/he would produce none at all. As a recent report from the College Board succinctly states: "the format of a test ... sends a strong signal ... about the types of thinking and learning that are valued in our society. A multiple-choice format carries a subtle message that recognising the right answer is more important than working out one's own solution, and that passive learning is sufficient ..." (The College Board 1990:3). Language test developers working in the structural-psychometric tradition would have been comfortable with the view of thinking and learning criticised in the quotation above, as, in

accordance with basic tenets of behaviourism[39], they drew little distinction in language testing between knowing and doing; recognising the right answer and producing it (working out one's own solution) would therefore not be perceived as being fundamentally different cognitive activities[40].

The third of Spolsky's suggested three phases in language test development is the 'psycholinguistic-sociolinguistic' phase. Morrow's (1981) characterisation of this phase as 'The Promised Land' is a fitting one, as it captures the emergence of language testing from a stifling preoccupation with form. However, as is argued below, even this 'promised land' had its weaknesses, a major flaw being its overemphasis on interpersonal, communicative aspects of language use. As Tannen (1985) suggests, it is possible to typify discourses as having an involvement focus or an information focus, and it is the unremitting concentration of this phase on the former of these foci that has contributed to challenges and modifications.

The emergence of new thinking in relation to language testing and, of course, language use, came about in response to a number of factors.

The increasing hegemony of English in dominant economies and the related increasing use of English in professional and business circles in non-English speaking countries such as India, China, and Japan, meant that ever larger numbers of second or foreign speakers of English needed at least some proficiency in English in order to be able to work effectively – even in their own countries. This need has recently become more urgent with the global influence of the internet, in whose use English dominates. Growth in the need to use English for professional purposes was reflected in the growth of what came to be known as the TESOL (Teaching of

[39] This claim is not to suggest that 'structuralists' support behaviourism in its fullest sense. What is being referred to here are the similarities between those aspects of behaviourism that favour atomistic approaches, which promote the learning of small bits of knowledge, or microskills, and then assume that these small bits can be amassed into coherent wholes.
[40] However, even at a commonsense level, it makes little sense not to distinguish between the ability to produce language and the ability to recognise appropriate forms, or even decode extended texts. Many non-German speaking students of philosophy or psychology, for example, need to be able to read seminal works in their field in the original – i.e. in German. After attending special language courses for this purpose, they may become sufficiently proficient in German to be able to read the relevant texts. However, it is highly unlikely that they would be able to produce meaningful extended prose in German themselves. That is to say, they will have developed a fairly high level of receptive skill in the target language, but will not have developed a similar level of productive skill. Assessing their receptive ability in German and then assuming a similar level of productive ability, as was the custom in the structuralist stage (and is still prevalent in many widely used tests such as the current TOEFL[40]), is thus highly likely to yield inaccurate and unhelpful estimates of candidate ability.

English to Speakers of Other Languages) industry, accompanied of course by certification and evaluation – i.e. assessment. The kinds of language assessments that would be useful to employers and workers would be those that reflect work-related language demands (i.e. work-sample, or the 'strong' form of language tests) rather than the kinds of assessments produced by the structuralists.

The inadequacies for these kinds of purposes of assessments based on structuralist views of language were also highlighted by the large and increasing numbers of foreign students who were applying for places in English-speaking higher education institutions and/or for employment in English-speaking countries. This increase led to an urgent need for the development of selection procedures that could relate to future performance in ways that tests designed along structural lines were not equipped to do. Higher Education institutions, for example, were not so much interested in whether an applicant could construct grammatically felicitous sentences, as in whether the applicant's grammatical proficiency would be likely to impact significantly on her/his academic performance once admitted to the institution.

In addition, the gap between developments in language teaching arising from cognitive and situative theories of learning and knowing, and the ways in which language was assessed, was widening. New understandings of human interaction and communication such as those reflected in the New Literacy Studies approaches discussed in Chapter Five on the one hand, and existing testing approaches on the other, made it essential for new assessment approaches to be developed. Changes in communicative technologies also played a part here. Goldman argues, for example, that a major impetus for the challenge to communicative competence as an adequate explanatory heuristic for understanding situated language learning and use, is the dramatic growth in systems for information storage and retrieval, which are "... redefining literacy requirements for the 21st century" (Goldman 1997:358). As he suggests, learners today are required to navigate through multiple sources of texts, which are frequently 'multiply linked', using a variety of media.

In this new phase (the psycholinguistic-sociolinguistic phase which became known as "the communicative competence movement"), the importance of context was fore-grounded, although,

as McNamara (1996) notes, no major theoretical advances were initially made in understanding the nature of abilities that underlie performance. The lack of progress in this regard has been a seriously constraining factor in the development particularly of tests which aim to predict future performance in specific contexts. This lack of progress has resulted in difficulties for test developers as they struggle to make connections between research on communicative competence and the development of frameworks that can be operationalised in the form of tests (Jamieson et al 2000). In the absence of clear and useable theory of this kind, many test developers have placed more emphasis on content than on construct validity as a basis for test construction. Some of the problems that arise in consequence of this emphasis are discussed below.

The theoretical underpinnings of the communicative competence movement were provided by Hymes (1967, 1972), who proposed a distinction between knowledge of a language and actual use of a language, as well as between knowing what could be called the rules of the language, and knowing how they should be used. Davies (1988) terms these knowing 'that' and knowing 'how', which is somewhat similar to the distinction drawn by Bialystok and Sharwood-Smith (1985) between 'knowledge' and 'control'. In addition, Hymes extended the notion of knowledge to include sociolinguistic knowledge as well as psycholinguistic knowledge. He coined the term 'communicative competence' to refer to this extended view of language.

In so doing, he set the stage for theorists working in applied linguistics and testing to move beyond the confining theories of language learning and use based on behaviourism, and divorced from contexts of real use. Perhaps the most influential, early discussion of communicative competence in relation to language testing was that contained in the work of Canale and Swain (1980). In brief, they proposed a model of knowledge of language as comprising the following kinds of competence: grammatical (e.g. knowledge of vocabulary, rules of word and sentence formation), sociolinguistic (e.g. appropriacy in terms of topic, purpose) and strategic (e.g. the use of coping strategies to improve output). As can be seen these relate to, respectively, linguistic, social and cognitive aspects or areas of knowledge.

Canale and Swain do not, however, as McNamara (1996) points out, address Hymes's elaboration of Chomsky's (1965) notion of performance. Both Chomsky and Hymes distinguished between actual use (e.g. correctly using the passive voice) and the potential or ability for actual use (e.g. knowing how to construct the passive form, and in what circumstances it would be appropriate and/or effective). Both would agree that knowledge of the passive voice (i.e. that the subject and object change places and so forth) is a part of competence. However, Chomsky included both of these: that is, (i) ability for use, which he rather confusingly terms pragmatic competence, and (ii) actual use, under performance, while Hymes conceptualised the latter (actual use) as constituting performance, and the former (ability or potential for use) as being part of competence. In other words, Hymes included, as part of 'competence', all those factors that are not part of an actual performance. The notion of an actual performance is perhaps more helpfully thought of as a 'product' in this sense`. Canale (1983a:6), too, distinguishes between "... underlying capacities ..., and their manifestation in concrete situations".

Hymes's theory of communicative competence is extremely helpful for language testing. In distinguishing between the roles played by 'ability for use' on the one hand, and knowledge of linguistic and sociolinguistic conventions, on the other, in actual use (such as the writing of a test), he drew attention to the role of

> "... a range of underlying language-relevant but not language-exclusive cognitive and affective factors (including general reasoning powers, emotional states and personality factors) which are involved in performance of communicative tasks" (McNamara 1996:59).

This distinction is particularly important for contexts where these 'language-relevant but not language-exclusive' factors are likely to be very different for different candidate groups, as is the case where educational disadvantage is known to exist. Researchers such as Garcia et al (1999), Steele (1997), and Steele and Aronson (1995) argue that many African-American students have internalised a negative stereotype which adversely affects their test performance in situations where they are in meaningful competition with their white student peers – a phenomenon known as 'stereotype threat'. Essentially, stereotype threat is "... a situational predicament – felt in situations where one can be judged by, treated in terms of, or self-fulfill, negative stereotypes about one's

group" (Spencer, Steele & Quinn 1999:6). In this view, the test performance of African-Americans would not reflect their knowledge of linguistic and sociolinguistic conventions so much as it would their ability to make this knowledge manifest in test performance. The relatively weak test performance of African-Americans on high-stakes tests would thus be perceived to be a consequence, at least in part, of internalised negative stereotyping.

Communicative competence, as proposed by Hymes (1972), requires us to "... attend to the entire package of a particular speaker's linguistic performance ..." (Hamp-Lyons & Kroll 1997:2). Some of the features of language use which the communicative competence movement, in response to this requirement, sought to incorporate into language testing are the following (Morrow 1981): the notion of interaction (for example, between individual and text, or another individual); the lack of predictability in real-time processing that occurs in communication; the role and impact of purpose in communication; and authenticity (of task and language use). This list can, as Alderson (1981b) points out, be expanded, but it serves nonetheless as a pointer for the kinds of abilities and attributes that language tests in the structural tradition failed to capture, but were paramount for communicative competence theorists.

It was suggested above that the Canale and Swain (1980) paper did not adequately deal with the issue of how language knowledge comes to be translated into performance. In attempting to address this limitation, Canale (1983a, 1983b) added a fourth type of language knowledge to the three proposed earlier. That is, to grammatical, sociolinguistic and strategic competences, he added discourse competence, which he conceptualised as concerning mastery of cohesion (the use of linguistic markers such as conjunctions to connect text) and coherence (the arrangement of ideas to enhance meaning) in discourse. This fourth type of competence, along with a broadened understanding of the role of strategic competence (which acts as an agent of use on the other three areas of competence in the enacting of performance), brought the notion of communicative competence more in line with that of Hymes's original conception. Nevertheless, Hymes's notion of ability for use is not elaborated, and as such Canale's framework is not particularly helpful for test development, particularly in regard to the 'strong' form.

Bachman (1990), too, conceives of strategic competence as a pivotal component in his exposition of the major interactions involved in language use. In summary, his model comprises three main components:

- topical knowledge structures. These include the information base constructed by an individual (e.g. cultural knowledge, content-related knowledge).
- language competence (knowledge of language). This component comprises organisational competence (further elaborated into grammatical and textual knowledge) and pragmatic competence (functional knowledge – e.g. heuristic or ideational functions - and sociolinguistic knowledge).
- strategic competence. Bachman (1990:102) defines the function of strategic competence as being "... to match the new information to be processed with relevant information that is available (including pre-suppositional and real world knowledge) and map this onto the maximally efficient use of existing language abilities". It is thus not so much an area of knowledge but, as McNamara (1996) suggests, is more usefully thought of as an ability – a capacity to synthesise and make manifest what is stored as knowledge and available as input. As with the Canale notion of strategic competence, however, Bachman's inclusion of strategic competence serves mainly as a marker that something important happens in the test-taking situation that cannot be explained simply in terms of knowledge, rather than providing clarity on the nature of its role.

The elaboration of the Bachman (1990) framework by Bachman and Palmer (1996) reconceptualises strategic competence as a group of metacognitive strategies. These strategies are identified as goal-setting, assessment, and planning, which all act with each of the other two cognitive components of the model: viz. topical knowledge and language knowledge, as in the 1990 model; and affective schemata. Affect (or affective schemata) appears as a component of language ability, on the grounds that, in combination with particular task characteristics, these schemata influence the ways in which individuals tackle tasks. Examples of such schemata are to be found in the phenomenon of stereotype threat, described above. Personality characteristics are

another illustrative area. Individuals characterised as 'novelty seekers' (Hamer & Copeland 1998:30) – that is, individuals who "... find pleasure in varied, new, and intense experiences ... [and] are willing to take risks for the reward of the new sensation ..." might react very differently to novel test formats, and such-like, than low novelty seekers, who tend to avoid risk if at all possible.

This new component (of affect) begins to address Hymes's 'ability for use' aspect of language ability by starting the long and slow process of explication. However, Bachman and Palmer simply advise test developers of its importance in the test-taking experience, and suggest that ways in which test anxiety can be avoided, and the best performances elicited, should be sought and used.

The TOEFL 2000 project at the Educational Testing Service provides an interesting illustration of the apparent failure of the communicative competence movement to provide a clear framework for test development. This project has as its aim the development of a new, computer-based TOEFL test that is more firmly based on a model of communicative language use than the existing test; is less reliant on multiple-choice formats; uses more constructed response tasks; integrates the language modalities (e.g. reading, writing, listening, speaking) wherever possible; and relates more directly to the academic environments in which candidates will find themselves. The project development team started work by conducting an extensive review of the literature on communicative competence, as well as by commissioning several papers on various constructs such as those by Hamp-Lyons and Kroll (1997) on writing, and Hudson (1996) on reading. The aim of these efforts was to develop a framework to use for the new test. However, when the results of all these initiatives were reviewed (Taylor, Eignor, Schedl & DeVincenzi 1995), it was reluctantly concluded that neither existing models of communicative competence, nor the papers commissioned by or written on behalf of the TOEFL 2000 Committee of Examiners, could provide a useful framework for test development.

This difficulty is by no means peculiar to the TOEFL development team, as is pointed out by McNamara (1996). One of the consequences, however, is that language test developers have fallen back on content specification as a basis for test development, at the expense, somewhat, of construct development. In essentially reducing or deconstructing a domain into a set of small,

discrete, observable behaviours, they are in danger of returning to behaviourism – as Messick (1994:17) notes:

> "... the preemptive emphasis on tasks and performances in the task-centred approach ... may not only bring behaviourism back into education by the rear door but, in effect, also behaviourism's talisman and shield, the operational definition".

In many ways, language tests in the 'strong' form as defined by McNamara (1996) are in danger of falling back into behaviourism. In the field of occupational training and personnel selection (Ryans & Frederiksen 1951), such performance tests are known as 'work-sample' tests. It is difficult (and unnecessary) to attempt to mask the strong similarity of work sample tests used in personnel selection to the kinds of tests used in adrnissions. However, the use of criterion situation tasks in test development need not mean that concerns about the nature of the abilities underlying task performance, including fundamental concerns about the nature of language proficiency, are simply ignored. Whether they are adequately dealt with or incorporated is the issue that needs analysis, and Messick's caution is necessary and timely.

Whatever the limitations, it is clear that tests within the communicative competence paradigm reflect a richer conception of what it is to learn and/or know a language, and of how this can be demonstrated, than did tests of previous eras. Most importantly, they are based on a notion of how individuals will perform in the 'real world' – that is, outside of the test situation. Such tests have come to be known as 'performance tests', defined by Fitzpatrick and Morrison (1971, cited in McNamara 1996:10), as those "... in which some criterion situation is simulated to a much greater degree that is represented by the usual paper-and-pencil test". The focus of interest is "... how well individuals can *do* something, as opposed to determining what they know *about* doing something" (ETS Trustees' Colloquy 1995:1, italics in original).

Nevertheless, few models of language proficiency deal adequately with questions relating to how abilities underlying performance can be inferred from a performance. Because of this, 'strong' language tests (McNamara 1996) – i.e. performance tests, which aim to predict performance in a defined (criterion) situation - are in a difficult position. They need to specify the criterion situation in

such a way that it can be adequately reflected by the test. However, without a clear understanding of how underlying abilities relate to each other, or of the role of non-verbal cognitive and affective factors in actual performance, it is extremely difficult to develop an understanding of the ways in which a task is experienced by a learner, and therefore of how the performance is constituted by the learner[41].

As Alderson et al (1995:226) suggest, "... the communicative revolution has become orthodoxy ...", and is now being challenged by views of language learning and use that stress the role of cognition and active processing of language-related data as fundamental. Claims of the importance of such aspects do not deny the roles played by sociocultural factors, but insist on a recognition of

[41] The following example illustrates how cognitive research on learning from text might influence the development of tests of academic literacy. For decades, candidates in language tests have been required to construct summaries of reading passages, in order to test their understanding of these passages. The importance of summarising ability in (for example) higher education studies is undeniable, as most effective expository writing, including writing in the sciences, places a premium on crisp, accurate, concise summaries of processes, theories, and arguments. Therefore, for an admissions test for Higher Education, including summarising as an important skill to be assessed is quite legitimate. The problem is with the way in which it is assessed. Generally, it is only the final product that is assessed, and the scoring is conducted on the basis of rating guides that look for numbers and organisation of main points, etc. - i.e. the focus is on the relationship between the text and the summary. A number of problems arise in connection with this (product assessment), two of which are discussed below.

First, researchers have long recognised that creating a representation of the text is not the same as constructing a representation of the text's referential situation. As Perfetti (1989) suggests, the difference between these two can be likened to focusing either on the meaning or the interpretation of a text. If the aim of the assessment is to check on a reader's understanding of a particular passage, then a straightforward representation-type summary is appropriate. If, however, the aim of an assessment is to check on the reader's ability to understand and interpret information in a given passage in relation to other information (elsewhere in the test, or in some cases using prior knowledge), then a representation-type summary is not sufficient. As the discussion in Chapter Five emphasised, cognitive and situative theories of learning lay great stress on the connections made by learners between concepts, pieces of information, skills, other learners, etc., and as such the text's referential situation would be an essential requirement. A number of ways of eliciting the referential situation of the text (i.e. the full interpretation, including meaning, in Perfetti's use of the terms) has been suggested by several researchers (Kintsch 1988, Chi, deLeeuw, Chiu & LaVancher 1994, for example). These ways include requiring evidence that the knowledge has been applied to new situations, constructing explanations (verbal or visual) that incorporate the new information, or carrying out a procedure which requires use of the new information.

A different kind of problem (with assessing only the final product of the summarisation process) is posed by Hidi and Anderson (1986), who suggest that three broad cognitive operations are involved in the summarisation process. These are: (i) the process of *selection*, involving not only the identification of main points in the text to be summarised; (ii) the *condensation* of material, involving the process called superordination - the construction of new, higher level, more general concepts which enables decisions on the relative importance of the selected ideas to be made, and for fundamental reorganisation of material to be undertaken – in the summarisation task this reorganisation is necessitated primarily by the need for condensation; and (iii) the *production* of 'new' text, involving the formulation of concise, coherent verbal representations through the "... further integration, combination and transformation of the original text propositions" (Yeld & Hartman 1992:47). It should be noted that the suggested cognitive operations involved in completion of these summarisation tasks have not been established empirically, but are hypothesised on the basis of relevant research (e.g. Hidi and Anderson 1986, Kintsch 1988, Chi et al, 1994). Clearly, focusing only on the final product of this process – that is, the summary itself – does not enable score-based inferences to be drawn about individuals' mastery of or difficulties with the various stages involved in the process. These inferences are important as they touch on possible reasons for poor (or excellent) performance. For example, candidates who have had few opportunities to construct text themselves - such as the learners of German cited above, who needed and acquired mainly reading skills in German, or learners in educational environments which have not prioritised writing (Kapp 2000) – might produce poor summaries not because of difficulties with (i) or (ii), but because of difficulties in converting meanings ('source input', as Kirkland and Saunders refer to it) into text.

cognition (such as learning styles and strategies, motivation, and personality) in language learning and use. The implications of this approach are discussed in some detail in Chapter Five.

It was argued above that a clearer understanding of the abilities underlying performance is a prerequisite for the development of a framework within which language testers can attempt to elicit some representation or evidence of these abilities. For example, language testers might seek information on such questions as the following: What cognitive operations (including strategy use) point to successful language use? Under what conditions are these operations and strategies successfully employed? How do readers create a hierarchy of ideas?

Most fundamentally, of course, a framework to be used in admissions testing must deal adequately with the relation between language proficiency and academic achievement.

6.3 Language Proficiency and Academic Achievement

Links between language and intelligence (however defined) have a long and controversial history. Theorists such as Oller (1981, 1979) and Oller and Perkins (1980), for example, argued that 'global language proficiency' accounts for much of the differences in performance noted on a number of measures of such different phenomena as personality and IQ (both verbal and non-verbal), as well as school-based achievement measures. This global language proficiency, they suggested, plays a central role in all facets of the learning process in schools. This insight provided support for the strong 'language across the curriculum' movement in the 1970s and '80s, in which teachers of all subjects were called on to become teachers of language.

By far the most significant component of the global dimension proposed by Oller (1981) was that of intelligence, operating as a kind of 'pragmatic expectancy grammar'. The details of Oller's theory are not relevant here, except to note that this emphasis on anticipation or prediction as a critical cognitive activity accords well with current theories of knowing and understanding, which place a premium on the organisation and use of prior knowledge in making sense of new information and experiences.

However, at its extreme, this argument (that, in many ways, language proficiency is intelligence) was developed into one which equated control over the surface features of language with cognitive ability – so that individuals speaking non-standard forms of a language were assumed to be cognitively deficient in some way. Labov (1973), in response to this kind of assumption, argued that the poor levels of performance on academic tasks often demonstrated by low-SES black children in the United States is the result of low teacher expectations (derived from the assumption that nonstandard dialect use is synonymous with cognitive deficiencies) rather than of low levels of the kinds of knowledge required by academic tests. Taken to its logical conclusion, the assumption against which Labov was arguing implies that the appropriate form of remediation for children who are performing poorly at school (and who are also frequently non-standard English speakers) was instruction in standard English. When this instruction failed to improve school performance to any significant extent, however, it became evident that the relationship between language proficiency and academic success was more complex than could be expressed through an understanding of language proficiency as a matter of form, or surface detail.

The link between cognitive factors, language proficiency and academic success is however strengthened by evidence from several studies, which suggests that at advanced levels of language proficiency,

> "... different aspects of proficiency seem to be differentially related ... . Cognitive variables appear to be ... strongly related to discourse aspects of proficiency and to written aspects of proficiency ..." (Harley et al 1990:25).

Somewhat similar findings are reported by Hamilton (1991), Lopes (1992) and Sheridan (1991), cited in McNamara (1996), in studies of native-speaker performance on tasks in IELTS reading tests. Their studies suggest that the variation in ability on these tasks is closely related to the educational level of the participants. In the Hamilton (1991) study, the highest performance (that is, the highest group mean) was achieved by the group with the highest school-leaving scores, who had taken an academically more demanding course in the Technical and Further Education (TAFE) system in Australia. The Lopes study replicated Hamilton's study, but used university graduates. Again, the academically stronger group (junior barristers, an academic elite, versus

post-graduate teacher training students and their lecturers) achieved a higher group mean. Both of these studies were conducted on an IELTS reading sub-test, and together with Sheridan's companion study investigating performance on an IELTS writing sub-test, provide support for the view that reading and writing tap general non-linguistic cognitive qualities.

Cummins (e.g. 2000,1984, 1980) proposes two different conceptions of language proficiency that are useful in this context. He was concerned to understand why students who appear fluent in a language frequently experience difficulties when that language is used as the medium of instruction. That is, he was interested in the abilities lying behind the successful deployment of language in academic settings, and in whether and in what ways these were different from the abilities underlying the use of language in non-school settings. It was in this context that he argued for language proficiency to be defined in a way that could be related to academic performance. In suggesting that academic success requires using and understanding language in context-reduced situations, he drew a distinction between the use of language in context-reduced as opposed to context-ernbedded situations (Cummins 2000, 1984, 1980, for example). Cummins's argument was that tests arising from communicative competence theories of language use would tend to tap only one dimension of the language abilities required to function effectively in formal schooling. He called this dimension basic interpersonal communicative skills (BICS), and contrasted it with that of cognitive academic language proficiency (CALP) which was intended to capture the kinds of language ability needed to function effectively in schooling. As Cummins and Swain (1986:151) point out, it is "... necessary to distinguish between the processing of language in informal everyday situations and the language processing required in most academic situations".

The situations being referred to here are the typically decontextualised ones found in formal schooling contexts, and some of the processing demands arise from the absence, in many academic situations, of the normal supports found in conversation (e.g. nods, interpolations, gestures), and the need to function as both audience and speaker (Bereiter & Scardamalia 1982). More specifically, decontextualised language use refers to "... language used in ways that eschew

reliance on shared social and physical context in favour of reliance on a context created through the language itself" (Snow, Cancino, De Temple & Schley 1991). It can be defined as requiring:

- "... the linguistic skills prerequisite to giving, deleting, and establishing relationships among the right bits of information" (Snow 1987:6), and control of "... the complex syntax necessary to integrate and explicate relations among bits of information, and maintaining cohesion and coherence" (op cit:7).
- proficiency in processing text irrespective of mode (i.e. spoken or written) or medium (e.g. books, journals, visual material, electronic forms) where meaning is supported by linguistic rather than paralinguistic cues (Tannen 1985, Cummins 1982, 1984, Wells 1981).
- communication, irrespective of mode or medium, where the emphasis is on the message rather than the act of communication (Tannen 1985, Wells 1981, Arena 1975).

The first of Cummins's two conceptions of language proficiency relates to the concept of communicative competence, which is firmly embedded in linguistic and sociolinguistic theory. In this view, it is the act of communication that is prioritised, rather than the message. The communicative competence movement, which arose in reaction to a view of language as the sum of numerous discrete elements, understandably fore-grounded interaction, context, and authenticity in human communication. The role of world knowledge, and other relatively non-linguistic cognitive factors was not, however, adequately addressed. It is this lack that the second of Cummins's proposals attempts to address, and in doing so, it draws on psychological theory for its insights, rather than linguistics or sociolinguistics. It is based on

> "...an analysis of the requirements of language tasks with respect to two dimensions: the degree to which the language task is supported by non-linguistic contextual cues, and the degree of cognitive effort involved in task performance " (Cummins & Swain 1986:205).

Language proficiency is thus conceptualised along two continua, (i) degree of contextual support and (ii) cognitive complexity. The use of two continua rather than one represents an attempt to avoid the oversimplifying effects of dichotomising constructs into two categories, as well as more adequately representing the two kinds of task demand characteristics identified by Cummins. It also explicitly links language proficiency and cognitive theories of knowing and learning.

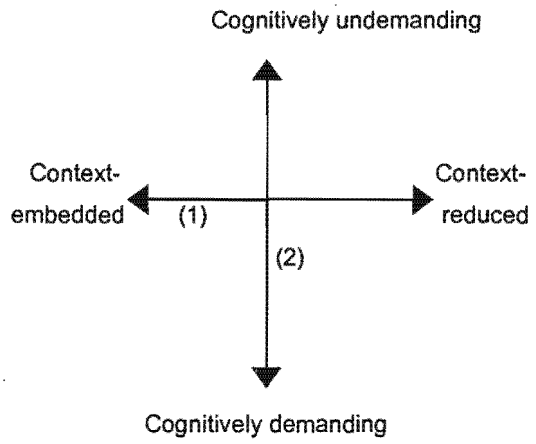

Figure 6.1: Range of Contextual Support and Degree of Cognitive Involvement in Communicative Activities (Cummins 2000, 1984, 1980)

Continuum (1) above relates to the degree of contextual support available for a task. Thus, by way of example, in a tutorial (or a conversation), cues and support are available through immediate reactions from an audience, and through real-time opportunities to actively negotiate meaning. This kind of situation is contrasted with activities (such as essay or report writing) in which precise and explicit use of language and linguistic devices is required to avoid misinterpretations that cannot immediately be corrected - these activities would be placed at the other end of the same continuum. It should be noted that the degree of contextual support varies for different individuals - for a painfully shy individual who feels both alienated from and threatened by a particular tutorial set-up, the tutorial event (the task) will not occupy the same place on the contextual support continuum than it would for a confident, high-achieving student.

The other continuum ((2) above) relates to the degree of cognitive effort required by tasks. As with the contextual support continuum above, a particular task or situation does not have a predetermined place on the continuum. The degree of effort (i.e. how difficult the task is for the individual) is more closely related, for an individual, to the individual's degree of mastery of the linguistic tools necessary for the task, than to some inherent quality of the task itself. These linguistic tools include, following the Bachman (1990) and Bachman and Palmer (1996) models discussed in section 6.4.2 below: (i) *topical knowledge* (e.g. the individual's knowledge, broadly defined, about the topic at hand); and (ii) *language knowledge* (which includes the following categories - grammatical, textual, functional and sociolinguistic). Interacting with and directing

these components are (iii) *strategic competence*, or *metacognitive strategy use* (the effectiveness of the individual at planning, monitoring, and modifying the language required and used by the test task), and (iv) the individual's *affective schemata*, which affect the way in which tasks are approached and undertaken.

Returning to Cummins's model: as the degree of mastery of the linguistic tools necessary for the task increases, the degree of cognitive effort required for the task decreases, and the same task assumes a different place on the continuum (it moves up towards the undemanding end)[42]. This, of course, begs the question of how mastery of the linguistic tools can be acquired: in this respect Cummins (2000:71) suggests that "... language and content will be acquired most successfully when students are challenged cognitively but provided with the contextual and linguistic support or scaffolds required for successful task completion".

It is with this second conception of language proficiency that connections with academic performance are often made. In particular, the linking of linguistic tools to task performance rather than to the more general notion of communicative competence makes a good deal of sense in contexts of widespread educational disadvantage. Poor educational systems do not, in general, provide appropriate opportunities for the development of task-related academic language skills. In this connection, the situation of educationally disadvantaged South African students is particularly difficult.

6.3.1 Academic Language and Educational Disadvantage

It is important to note that Cummins's framework was developed in a very different context to that of South Africa. It assumes that learners will have task-relevant linguistic skills already developed in their first language. This helps to explain the close relationship hypothesised by Cummins between the level of cognitive demand, or challenge, experienced by an individual and the level of linguistic skill possessed by that individual. That is to say, the Canadian students described by Cummins (1984), and Harley et al (1990) can, by and large, be assumed to be able to perform

[42] By way of illustration: Sophia, with fewer of the linguistic tools required by a particular task at her command than Sipho has, will find that task more cognitively demanding than he (Sipho) will. As Sophia's repertoire of task-related language

various language-related academic tasks in their first languages: it is in performing these same tasks in a language other than their first that the cognitive challenge is experienced. Educationally disadvantaged South African students, however, cannot be assumed to have developed these skills in their first languages. In the first instance, they are highly unlikely to have experienced conducive learning conditions for the development of relevant CALP skills in their first language. As the Report from the Ministerial Committee into the Senior Certificate (DoE 1998) states:

> "... there is evidence that a large proportion of our schools do not give students enough practice in reading – that is to say, in developing critical, selective, analytical and interpretive reading skills – and writing – in developing critical, creative, interpretive, eflective, analytical and transactional writing skills. This lack of opportunity for practice appears to be particularly prevalent in the teaching of African languages" (op cit:12)[43].

Citing evidence from examiners' reports, the report goes on to suggest that "... the teaching of African First Languages is aimed at language maintenance and language development, rather than at cognitive development as is the case with English First Language" (1998:42). Since the overwhelming majority of the candidates taking the Senior Certificate take an African Language at First Language level, this focus is clearly a serious problem.

Second, compounding this difficulty are the poor general conditions these students will have experienced in the learning of English (their second, or additional, language). Major factors contributing to these poor general conditions are: the generally low teachers' level of proficiency in English, the high reliance on rote learning and passivity of learners, the unavailability of appropriate learning materials such as textbooks, and the widespread chaos in the school system (Kapp 2000a&b, Plüddemann et al 1999, Vinjevold 1999, Macdonald 1990). In addition, the generally poor and counter-productive ways in which English Second Language is assessed at school and in the Senior Certificate exacerbates the problem (DoE 1998, Hansen 1997).

Third, for the majority of learners, the language of learning (the medium of instruction), which is overwhelmingly English, is not the students' first language. As Plüddemann et al (1999) comment

---

skills increases, however, the task will become less cognitively demanding for her, and she will move closer to Sipho's position on the continuum.

[43] Cummins and Swain (1986:94) highlight the implications of this as follows "... students' L1 cognitive/academic skills are just as important as L2 exposure for the development of cognitive/academic skills in L2".

in respect of the situation of Xhosa-speaking learners (in an English medium of instruction context and with an English first language teacher), "[I]n a situation in which the teacher understands perhaps half a dozen words or phrases in Xhosa, and the learner knows only enough English or Afrikaans to follow the basic instructions and to answer in monosyllables, interaction between teacher and learner is necessarily stunted" (cited in Vinjevold 1999:221). It is not necessary to spell out the implications of this situation for the learning environment. Mediated learning experiences (i.e. where experiences are selected, structured and interpreted for the learner by another individual or some other agency) are not simply important for the particular learning situation in which they are employed. On the contrary, they are crucial in enabling a child to learn in other situations. An absence or very low number of mediated learning experiences results in severely restricted learning opportunities which can have long-term consequences for the individual.

Despite these important contextual differences between the context in which Cummins developed his framework and the South African situation described above, the two-continua model proposed by Cummins provides a useful conceptual heuristic for test development. For example, in incorporating a developmental, dynamic component, it reminds test developers of the many variables that can impact on performance, and of the ways in which underlying abilities can change (or develop) and in turn impact performance differently, thereby according well with Vygotsky's notion of the ZPD. In concrete terms, it calls for attempts to be made to provide as much contextual support as possible, as well as to provide opportunities for acquiring as many of the relevant linguistic, task-related tools as possible. Most importantly, the model embodies a claim that "... academic performance and ability to perform more cognitively demanding, context-reduced language tasks are positively related" (Cummins & Swain 1986:205) and, in so doing, facilitates meaningful test validation studies.

6.4 Developing a Construct for a Test of Academic Literacy

It is clear that the kinds of language task relevant to Higher Education need careful consideration, so that these 'cognitively demanding, context-reduced language tasks' may be identified. Sections

6.4.1 and 6.4.2 below provide an overview of such an undertaking, restricted, however, to a consideration of language needs related to reading and writing in Higher Education. While all four language skills (reading, writing, speaking and listening) are important in educational contexts, the discussion here focuses only on reading and writing. There are a number of reasons for this restricted focus. Reading and writing are (arguably at least) the most fundamental in the learning activities of students. In addition, however, the requirements of cost effectiveness and comparability unavoidably restrict large-scale tests in the context of this study.

6.4.1 Reading and Writing in Academic Contexts

In undergraduate education in particular, the overarching purpose for *reading* is comprehension. Of course, students use language, and read, for many other purposes during their undergraduate years. However, the main purpose students read in the course of their degree studies is comprehension and learning. Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl (2000) subdivide this broad construct of reading into four kinds of purpose, each of which deals with a different aspect of the construct of "reading comprehension". These purposes are described below. It should be noted that they are hierarchically organised – a reader who is able to read for the purposes of basic comprehension is assumed to be able to read to learn, and so forth.

- Reading to find specific information. In this kind of reading purpose, readers are seeking to locate particular bits of information (e.g. by skimming), and to comprehend these. For example, a student might need to read about a particular experiment referred to by a lecturer in a tutorial. Examples of particular skills required in this task are word recognition and macro text layout comprehension.
- Reading for basic comprehension. This term refers to understanding the main gist of a text (the theme and the main points). It does not involve an understanding or comprehension of how the arguments support the main ideas, and so forth.
- Reading to learn. This purpose involves the conceptual integration of the information in a text. Key activities in this kind of reading include seeing and/or making connections, integrating information (for example, across arguments), organising information, and understanding rhetorical intent (such as cause/effect, contrast/comparison schemas of information

presentation). In essence, reading to learn requires the ability to stand apart from text, to demonstrate a critical stance.

- Reading to integrate information across multiple texts. The major difference between this kind of reading and reading to learn is that it requires readers to be able to make connections across texts, without there necessarily being obvious links within the individual texts. Readers therefore have to construct their own organising frames (Enright et al 2000). Typical frames of this kind are contrast/comparison, or the construction of tables to compare data.

It is important to note that these four kinds of reading do not directly correspond to a scale of task difficulty. That is, reading to integrate across multiple texts is not necessarily a more difficult task than reading to learn. If a text is very short, and/or very straightforward, making links between it and another simple text is a relatively easy task, whereas creating a concept map of a complex and abstract text can be extremely difficult. In developing items that tap these purposes for reading, it is important to keep item difficulty in mind.

All of these purposes can be described as ideational and/or heuristic uses of language, defined by Halliday (1973) as primarily to express propositions or to exchange or structure information. It is usually referred to in connection with the formal education setting, where much language use is undertaken in the pursuance of learning and for demonstrating learning as distinct from interactive communication.

Turning now to academic *writing*, it can be argued that in very general terms, undergraduate students write to fulfil two main purposes[44]:

- To transmit information (writing to display knowledge). Examples of this kind of writing abound in formal education, such as in examinations, and assignments. In this case the transmission is from one individual to another.

[44] Other purposes, such as using writing as a "channel for the human imagination" (Cumming et al 2000:3) are less widespread in formal education.

- To create knowledge (writing to learn). This purpose includes rewriting notes – for example in the form of 'mind-maps', where the information is transformed into another representational form and understood more thoroughly – as well as the drafting process which most writers undergo when engaged in a writing task.

Each of these purposes draws on the same kinds of abilities as the other language modes (speaking, listening and reading), although in writing, these abilities take form through "... heavily conventionalised scripts and text forms" such as laboratory reports, book reviews, and so forth. Both reading and writing require the participant (reader or writer) to operate in a highly context-reduced situation. In writing, the language user has the added requirement of ensuring that her/his intended meaning is communicated clearly in the absence of any other aids (the aim is to encode, not to decode). The writer is thus required to act as audience as well as actor, which requires considerable time, motivation, and cognitive effort. It is no wonder that one frequently hears of "writer's block", not "reader's block"!

One of the most striking characteristics of academic writing is that it entails a process. Theorists such as Raimes (1991), using a process approach, describe a series of stages: from the initial brainstorming to drafting, obtaining feedback, revising and editing, to submitting for publication or assessment. These stages are not commonly undergone by an individual writing a letter to a friend, for example, but are commonly undergone, to a greater or lesser extent, in academic writing.

A second characteristic of the kind of writing that takes place in academic contexts is that considerable planning is required. Of course, even writing a letter to a friend requires some planning, but it does not usually entail the drawing up of a plan, often with divisions and subdivisions, and the categorisation of material into main and supporting points.

Writing in academic contexts is, like other forms of writing, almost always a social undertaking, in that there is an intended audience. What makes academic writing so distinctive, for undergraduate students in particular, is the power relationship that exists between the writer (the student) and the

reader (the lecturer, examiner, tutorial audience). As most writing is undertaken for evaluative purposes, it is a somewhat stressful undertaking for students. Collaborative writing, the most social kind of writing, is less frequently used in academic contexts, although its use is expanding with increasing use of group project work.

A fourth characteristic of academic writing is that it is usually of a predetermined, specified length. Most writing assignments carry an explicit length requirement. Others have clear conventions - for example, a laboratory report follows a certain pattern. This puts a further constraint on the writer. Not only must s/he 'get things right', and 'get things clear' (that is, be clear about the topic and be clear in the writing of it) but s/he must do so in a space determined by someone else, or by some superordinate convention. These constraints do not necessarily make the writing task more difficult, but they do impose limits on what is acceptable.

6.4.2 Models of Language Ability

The Bachman and Palmer (1996) model of language ability occupies a central place in the field of second language assessment (Cummins 2000) and provides a useful framework for defining the notion of linguistic tools as understood in the Cummins approach. To recapitulate briefly, the Bachman and Palmer model conceptualises communicative language ability as comprising four components: (i) topical knowledge, (ii) language knowledge, (iii) metacognitive strategy use, and (iv) affective schemata. These components are elaborated below with particular reference to reading and writing in academic contexts.

Topical Knowledge

In the Bachman and Palmer model (1996), topical knowledge refers to the information (about the communicative situation and the subject matter) brought to a situation (e.g. a test, a conversation, a task) by an individual. As they point out (op cit:120), topical knowledge has long been regarded by test developers as a "... potential source of test bias ...", to be minimised whenever possible. However, for tests that aim to predict future performance in a situation in which language will be one of many variables, the role of topical knowledge ought not to be downplayed. On the contrary, as has been frequently stressed in this and previous chapters in this study, the role of expertise in

cognitive functioning is an important one, and assessing what students know and can do with language is considered to be highly dependent on the availability and use of topical knowledge.

Test developers have several options in relation to topical knowledge. In one of these options, it could be regarded as forming part of the construct definition. An example here would be if doctors needed to be certified as competent to practice their profession in an English-medium environment. In such a case the test might legitimately assume a considerable degree of core medical knowledge, and it would not be considered irrelevant to include topical knowledge. It is possible that in assessing the doctor's performance, an attempt would be made to separate the knowledge of the doctor about (for example) malaria or food poisoning from her/his effectiveness as a communicator. Nevertheless, the topic would be viewed as an integral part of the assessment aim, and the final result would incorporate both aspects. In this case the test construct clearly includes topical knowledge (and in doing so, as McNamara (1996) points out, raises serious questions about the ethics of forcing foreign doctors to pass what amounts to an oral medical examination while indigenous doctors are exempt from having to do so).

In another option, topical knowledge could still be regarded as essential for the elicitation of CALP skills. However, if the candidates for whom the test is designed come from a wide variety of backgrounds, topical knowledge would be more likely to become a source of test bias rather than a fertile and useful source of information on which tasks could be based. In this case, an appropriate response would be for the test developers to provide appropriate opportunities for the desired topical knowledge to be developed. To continue with the example above, the test could contain a variety of texts (e.g. illustrations, articles, brochures, tables) on malaria or food poisoning, and candidates could then be required to discourse (verbally or in written form) on the topic. In this sense it could be said that although topical knowledge is included in the test construct, it is not assumed that learners bring it with them to the test. The challenge for test development in this option is to provide the kinds of information that can act as topical knowledge.

Language Knowledge

In the Bachman and Palmer (1996) model of language ability, language knowledge (the second of the four components listed above) comprises two broad categories, namely organisational and pragmatic knowledge. Together, these make up "... a domain of information in memory that is available for use by [the] metacognitive strategies in creating and interpreting discourse in language use" (op cit: 67).

Organisational knowledge is concerned mainly with formal structures of language, where the emphasis is on accuracy and control. It includes grammatical and textual knowledge, as detailed in Figure 6.2 below. Pragmatic knowledge, on the other hand, is concerned with the relationship of these formal structures to meaning, and to the ways in which particular settings and purposes impact on the structures and their meanings. These two broad sub-domains (of language knowledge) are explicated in Figure 6.2 below, which contains a representation of the Bachman and Palmer (1996) model of language knowledge.

It is clear that not all of these categories will be equally useful for all purposes. For example, morphology, listed under Bachman and Palmer's 'grammatical' category, is unlikely to be as critical for performance in a tertiary academic environment (and therefore in an admissions test) as it would be in a test for language translators. The most important categories of language knowledge in an academic context are those of ideational and heuristic knowledge – that is, those aspects of language knowledge that enable a student to understand and use language to express ideas and relationships between ideas, and for purposes of learning.

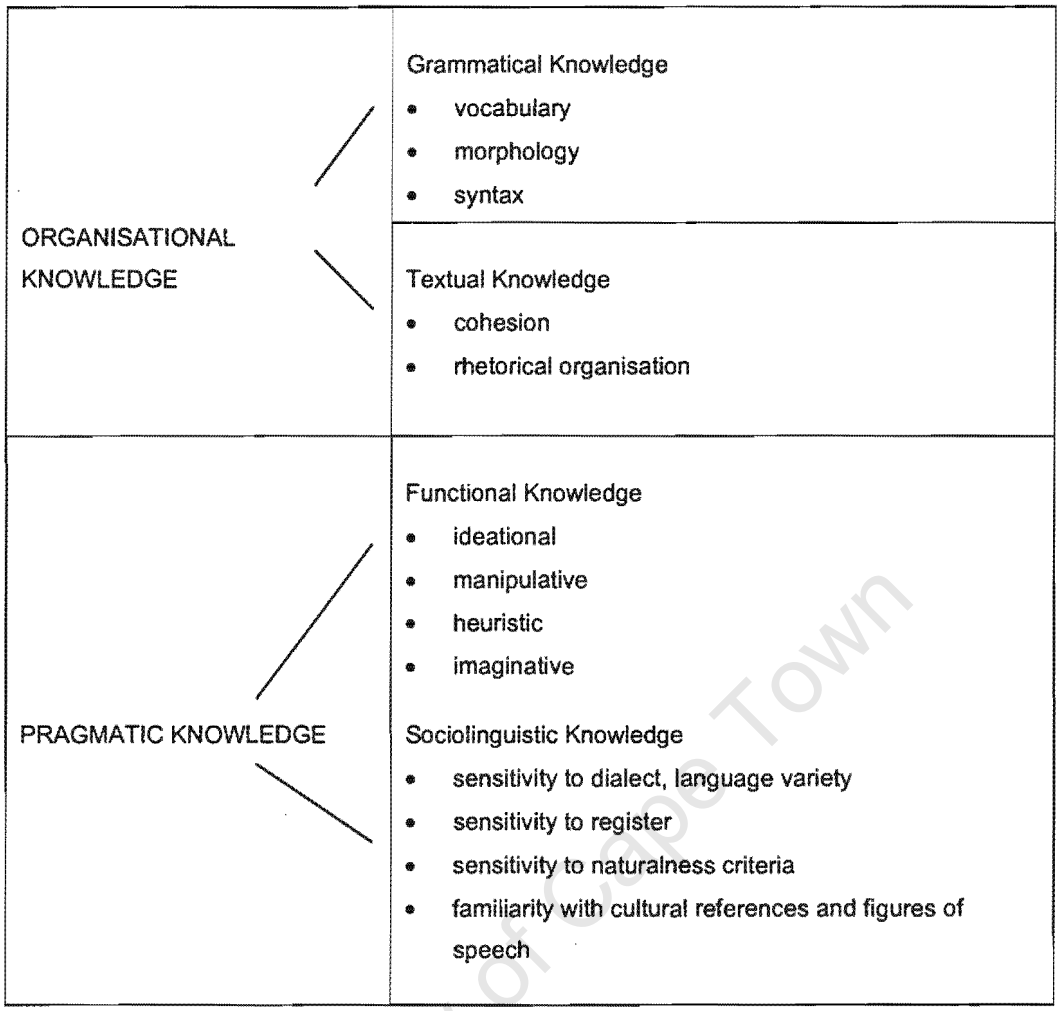

Figure 6.2: Language Knowledge (Bachman & Palmer 1996)

Strategic Competence (Metacognitive Strategy Use)

In discussing the learning problems experienced by educationally disadvantaged students, Moll and Slonimsky (1989:165) suggest that because these students have not had exposure to effective and varied approaches to learning in a formal context, they "... have never been required to mobilise and coordinate their existing cognitive operations in the development of academic skills ...". This 'mobilisation and coordination' is identified as a crucial component in effective learning. In the Bachman and Palmer model, the 'higher order executive processes' referred to by Moll and Slonimsky are conceptualised as a set of metacognitive strategies. In particular, goal-setting, assessment and planning are singled out as essential strategies.

It seems clear that the skillful and appropriate deployment of metacognitive strategies plays an important role in effective language use (e.g. Cohen 1998, Purpura 1996 – cited in Bachman &

Cohen 1998). Testing metacognitive strategy use, however, presents many challenges. In most cases in which the use of metacognitive strategies has been examined, the research has relied on questionnaires, think-aloud protocols (during or after the test writing event), or interviews of some kind. For a high stakes test such as an admissions test, these approaches are clearly neither realistic nor desirable. 'Politically correct' answers can, for example, be learned (McNamara 1996). Nevertheless, it could be argued that providing the kinds of conditions within a test that promote or activate the use of metacognitive strategies is likely to provide, at the same time, the kinds of support put forward in the Cummins quadrant model as essential to mediate cognitive demand.

Affective Schemata

Test takers' performance on tasks can be influenced by many factors other than those intended by the test developers. An obvious example is the inclusion of emotive topics – in the South African context, such topics could include apartheid, racism, AIDS, poverty, abortion, and so on. Candidates' reactions to these topics would in all likelihood impact on their ability (or willingness) to engage with the tasks as planned by the developers.

Another example can be seen in the existence of 'stereotype threat', discussed in 6.2 above. Test taking, which must be regarded as a particularly acute 'situational predicament', is highly likely to invoke such stereotypes for affected groups, such as educationally disadvantaged students, or women in numeracy testing initiatives.

Sections 6.3 and 6.4 in this chapter have highlighted and analysed important facets of language use in an academic context. In 6.4.3 below, the implications of this analysis are brought to bear on the development and articulation of a construct for an academic literacy test in a context of extreme educational disparities in provision.

6.4.3 Towards an Appropriate Construct for a Selection Test for South African Higher Education

In Chapter One, section 1.2, the central research questions in this study were identified. The first of these was: *On what basis ought a selection test to Higher Education, in a context of widespread*

educational disadvantage, to be constructed? In other words, what would be an appropriate construct for such a test?

In addressing this question, the following account of the construct of language proficiency on which such a test should be based, is outlined. The construct draws on the discussion in the first six chapters of the study, and lays the foundation for the analytical part of the study that follows. The construct is grounded in theories of dynamic assessment, theories of knowing and learning, language testing theory (including notions of language proficiency), and understandings of typical academic tasks, based largely on inputs from expert panels.

The construct states that the tests will:

- aim to predict the performance of candidates in future situations in which language will be an important, but not sole, variable. They will be based therefore on a notion of language-as-vehicle rather than language-as-target (in McNamara's 1996 terms, the tests are 'strong', not 'weak', language tests).
- acknowledge the effects on cognitive functioning of the quantity and quality of prior mediated learning opportunities experienced by an individual, and attempt to develop and include strategies to address negative effects where possible. The construct thus embraces both Vygotsky's notion of the ZPD and Cummins's notions of the interrelationship of contextual support and cognitive demand in the development and deployment of CALP skills. The implication of this is that dynamic approaches to assessment will need to be seriously considered and incorporated as far as is feasible. The inclusion of this requirement or aspect of the construct is, as far as is known, unique in a pencil-and-paper selection testing context.
- be based on a notion of knowing and learning which views learners as actively involved as individuals and in collaboration with others, in creating and negotiating meaning in a wide variety of settings. This process of conceptual development is seen as highly dependent on specific areas of expertise involving knowledge and information, and on the connections between these.

- be based on a componential model of language ability, which comprises topical knowledge and language knowledge, mediated by strategic competence (metacognitive strategy use) and affective schemata (see 6.4.2 above). In respect of topical knowledge, this will be included as an important component of the construct definition. This knowledge, however, will be limited as far as is possible to information provided in the test itself. Prior knowledge of the topic of the test will not be included in the construct definition. In respect of language knowledge, this will be understood as comprising the following categories or kinds of knowledge: organisational knowledge (grammatical and textual) and pragmatic knowledge (functional and sociolinguistic). The various components of these are listed in Figure 6.2 above.
- not directly assess metacognitive strategy use, but will aim to ensure its inclusion through incorporating tasks requiring deployment of such skills as planning and assessment/evaluation of alternatives.
- acknowledge (and attempt to mitigate) the impact of affective schemata on performance.
- be based on a coherent understanding of academic tasks.
- use a mixture of item formats in an attempt to achieve as comprehensive a range of representations of candidates' knowledge and skill as possible.

6.7 Conclusion

In this chapter, a brief history of language testing has been outlined, alongside changing views of what it is to know and use language. The increasing importance accorded to 'ability for use' was discussed, as well as the challenges this poses for the assessment of language proficiency. In addition, the different emphases and orientations of 'strong' (work-sample) and 'weak' language tests were sketched, and it was concluded that academic literacy tests developed to be used as admissions tests fell into the former category. The impact of educational disadvantage on the development of CALP-type language skills was explored in Chapter Five and returned to in Chapter Six.

The chapter concluded by articulating the basis on which it has been argued an appropriate academic literacy test could be developed for the South African context. The construct draws on

the theories and arguments outlined in preceding chapters, and prepares the ground for the validation of the PTEEP tests in the following chapters. These tests, developed to meet the needs of the University of Cape Town to change the composition of its student body as rapidly as possible, as well as the demands of black students for increased access, have not been comprehensively validated against a detailed set of specifications. It is this task that is undertaken in Chapters Seven to Twelve, with the aim of addressing the second major research question, viz.

To what extent are the PTEEP tests, developed to identify talented but educationally disadvantaged candidates whose SC results would not necessarily reveal their abilities, valid in terms of the construct articulated above (section 6.4.3)?

CHAPTER SEVEN
THE PLACEMENT TEST IN ENGLISH FOR EDUCATIONAL PURPOSES: ORIGINS AND HISTORY

7.1 Introduction

Chapters One to Six of this study covered several important and relevant issues relating to assessment, such as fairness and bias in testing; positive and negative aspects of large-scale assessment; theories of knowing and learning and their implications for assessment; academic literacy and related assessment approaches; and dynamic assessment and educational disadvantage. These chapters concluded by specifying certain requirements that must be met, in the South African context of widespread educational disadvantage coupled with a high degree of inequity in educational provision, by admissions tests in the area of academic literacy. In so doing, the ground was prepared for the comprehensive validation investigation contained in subsequent chapters.

In Chapters Seven to Twelve, the academic literacy tests developed by the Alternative Admissions Research Project - the English for Educational Purposes Placement Test (PTEEP) tests - are subjected to a validation process. In this process, their claims of being grounded in principles of good assessment practices, sound and appropriate theories of knowing and learning as applied to assessment possibilities, and of being responsive to and appropriate in a context of widespread educational disadvantage, are assessed.

The requirements of good testing practice can be divided into two broad categories. The first of these categories comprises what Alderson et al (1995) suggest could be termed technical concerns, relating to such matters as establishing validity and reliability, the test development process, scoring and rating issues, and the production of manuals and user guides. The second set of requirements can be understood as relating to professional concerns, such as appropriate test use, and test fairness. Both of these sets of concerns are addressed in Chapters Seven to Twelve.

The validation study is structured as follows. In this chapter, the origins and history of the Alternative Admissions Research Project, in which the PTEEP tests play a central role, are described. The aim of the discussion is to clarify the particular test development trajectory taken by the PTEEP. The original role of the Project, viz. to be a key means of recruiting and selecting black students for the institution, as opposed to being established as an assessment enterprise, is spelt out, as this origin has had serious consequences for the ways in which data gathering and testing procedures generally have been conducted. That is, the chapter highlights the Project's beginnings as an advocacy project for the admission of black students to what was substantially a white institution, and clarifies how this shaped and constrained its development as an assessment project. In addition, certain specific challenges arising directly from its context are discussed.

Chapters Eight and Nine focus on internal aspects of validity. Chapter Eight investigates the construct validity of the PTEEP. It ties together the notion of academic literacy put forward in earlier chapters with the particular purposes and context of the PTEEP tests. Chapter Nine analyses content, face and response validity. That is to say, it assesses the extent to which the test can be said to embody the test construct (content validity) and the extent to which it is perceived and experienced as valid by test takers and users (face and response validity).

Chapters Ten and Eleven address external aspects of validity. That is, they focus on the relationship of test performance to phenomena external to the tests, such as future academic performance, and/or performance on other supposedly similar tests. In Chapter Ten, predictive validity - a particularly crucial form of validity for admissions tests - is discussed. Chapter Eleven

focuses on concurrent and consequential validities, assessing the extent to which the tests differ from or are similar to existing tests, and their impact on the context into which they have been inserted.

Chapter Twelve contains an analysis of the reliability of the PTEEP tests.

Where applicable, the analysis of the PTEEP is supported by empirical investigation. Where such investigation is not possible, recommendations are made on how the lack of empirical data could be rectified in future.

7.2 Origins and History of the PTEEP Tests

In the late 1950s and early 1960s, universities in South Africa faced many restrictions in terms of whom they could admit. In 1953, the notorious Bantu Education Act was passed, which restructured schooling along racial and ethnic lines, and in 1959 the inappropriately named Extension of University Education Act extended this restructuring to higher education. This ensured that there was a mechanism for excluding students on racial grounds, in line with the increasingly pervasive system of apartheid. In consequence, from 1959 to the late sixties, there was a considerable decrease in the numbers of black (particularly African) students at residential universities[45].

In the 1970s, however, black enrolments began to increase at the English-medium institutions, whose administrations went to great lengths to obtain permits for prospective students. Permits were usually only granted if students registered for certain courses (such as Italian, or Biochemistry) which were not available at the few institutions set aside for black students, such as the then University College of Fort Hare. This permit system, while inimical to academic planning, was fairly easily manipulated by students and institutions, as once they had been admitted, they could transfer to other courses (usually in their second year) and then continue studying while the institution embarked on a series of lengthy appeals. In the 1980s, the State established several new universities in the so-called independent 'homeland' areas – such as Bophuthatswana,

[45] These numbers, however, had never been high (e.g. Bunting 1994).

Transkei, and Venda. These new institutions were intended to accommodate the needs and aspirations of the great majority of black students, of whom only a very small number would be allowed to register at the White institutions. This small number would be admitted via the proposed new regulatory mechanism of quotas, which would replace the old permit system. In essence, this would force the institutions themselves to assume responsibility for excluding black students. Public outrage at this proposal ensured that the quota system was not implemented, although it remained, somewhat threateningly, on the statute books until 1994.

Until the late 1980s, then, institutions could, to some extent, manipulate the permit system, and admit students on merit. However, the Committee of University Principals, comprising the vice-chancellors of the Historically White Universities (HWUs), alarmed at high and escalating failure rates, and very low throughput rates generally at all levels of the system, recommended at this time that universities raise the level of their admissions criteria. This recommendation, coupled with the already very low levels of performance in the DET Senior Certificate examinations and the growing crisis in black schools, meant that the great majority of places at white universities continued to be filled by white applicants. By way of illustration, in 1983, 95% of the DET students who obtained a matriculation exemption (the minimum requirement for eligibility for degree study) attained D or E aggregates (Badsha, Williams & Yeld 1987). This was, by and large, below the aggregate level required by the white institutions. In addition, the particularly appalling state of mathematics and science education in DET schools meant that even if DET students attained a high enough aggregate to be considered, they would be unlikely to achieve the kinds of results in these subjects to be considered eligible for admission to any but Humanities faculties.

In addition, the widespread suspicion that DET Senior Certificate (SC) results did not effectively predict future academic performance was supported by a growing body of evidence (Badsha et al 1986, Shochet 1986, Potter & Jamotte 1985). The universities were therefore faced with the highly unsatisfactory situation of having to select students on a basis known to be unreliable. To make matters worse, the DET examinations produced a very restricted range of scores, making selection on the basis of a rank well nigh impossible, even if the scores were reliable. The consequence of

this lack of predictive validity was that many students who could have had the potential to succeed were denied access, and the universities were thus the poorer.

Table 7.1 below shows the number of students in 1988, classified by population group, at all residential universities in South Africa[46]. It displays the very low numbers of black students in the system as a whole, and particularly at the HWUs[47], at around the time (1987) that the AARP project was established at UCT.

Students	English-medium HWUs	% of enrolment at English-medium HWUs	Afrikaans-medium HWUs	% of enrolment at Afrikaans-medium HWUs	All universities	% of enrolment at all universities[48]
Black (African)	4,759	10 %	673	1 %	90,345	32 %
Coloured	2,384	5 %	1,497	2 %	18,166	6 %
Indian	3,969	8 %	91	0 %	19,048	7 %
White	36,655	77 %	65,844	97 %	155,764	55 %
Total	47,767	100 %	68,105	100 %	283,277	100 %

Table 7.1: Student Headcount Enrolments by Race at South African Historically Advantaged Universities (1988) [Adapted from: Cooper & Subotsky 2001]

By 1988, a number of Academic Support units had been established at English-medium HAUs in particular. They tended to target the few black, educationally disadvantaged students at those institutions, and to offer remedial or foundational work in core disciplines. The orientation towards a 'quick-fix' of a minority rather than a more systemic response was based on an assumption that considerably underestimated the magnitude of the educational problem created by Bantu Education. An example of this underestimation can be seen in the early naïve assumptions surrounding AARP which, supported by donor funding, was conceived of as a five-year project whose primary objective was to increase the numbers of black students at the institution through:

> "... devis[ing] selection criteria to identify educationally disadvantaged students with the potential to succeed at UCT, given the academic support presently available. Secondary objectives include ... establishing the legitimacy of the project in the eyes of the communities it is intended to serve, and determining the level of logistical support necessary in such an endeavour" (Badsha, Williams & Yeld: 1987).

[46] That is, not including the distance-learning institutions, the University of South Africa (UNISA) and Vista University.
[47] It also shows that the only real access opportunities for black students to HAUs were to the four English-medium institutions.
[48] African students comprised over 97% of all enrolments at the Historically Disadvantaged Institutions.

In other words, the institution aimed to increase the numbers of black students at the institution while minimising risk in terms of academic performance. Various approaches to selection could have been adopted at this stage[49]. Increasing the numbers of black students by increasing the size of the student body is one example of a possible response. UCT, however, had decided in the 1980s to limit undergraduate growth. Clearly, then, growth in black student numbers would entail the replacement of white students with black students. Other English-medium HWUs, in contrast, opted for increased growth as a way of accommodating growth in black student numbers. The decision by UCT in part explains its determination to ensure, by whatever means it could, that the black students it admitted would be successful. The implications of this emphasis on success for the new Academic Support Programme and the admissions project were serious.

Another possible response would have been to simply lower admissions requirements. A number of problems arise here. First, lowering or changing admissions requirements, unless carefully targeted, could lead, as Cloete and Pillay (1987) point out, to the admission of an increased number of white students with mediocre SC results. In a report on an Arts Faculty initiative at the University of the Witwatersrand, Cloete and Pillay (op cit) note that when a battery of tests was administered to a group of 670 students who had not met the faculty's automatic entry points requirements, 51% of the white applicants gained admission, compared to only 20% of the black applicants. Clearly, the black applicant group suffered here from adverse impact[50].

Second, the SC results of the great majority of black students writing under the authority of the DET system who succeeded in obtaining a matriculation exemption fell within a very restricted range, and lowering the admissions SC points requirements would in all probability simply have made selection more difficult. In 1993, for example, 93% of the 370,834 DET candidates obtained D or E aggregates. That is, 344,875 students obtained results within a 20% range. This range attenuation makes selection extremely difficult, as there is little to distinguish between applicants if one uses aggregate scores alone. The very narrow range of subject choice of DET students in particular, on top of the narrow range of scores, complicates the selection challenge even further.

[49] These options are discussed in some detail in Chapter Two.
[50] See Chapter Two, Section 2.7.1.

In 1992, for example, 98% of the students in the DET system took the vernacular language at the Higher Grade level (HG) as well as ESL-HG, 95% took Afrikaans Second Language HG, 66% had Biology HG/SG and 95% had History, Biblical Studies and Geography as subjects (Lötter 1994:22).

Third, the academic progress of educationally disadvantaged students, already a cause for concern, would, it was argued, in all likelihood worsen if admissions requirements were lowered. This concern co-existed uneasily with the fourth of the major problems associated with increasing access for black students by lowering admissions requirements: namely, that the DET SC results in particular were not reliable predictors of academic performance. Despite the apparent contradictions between these two positions (that the DET results were unreliable, yet lowering points requirements using these unreliable DET results would lead to the admission of weaker students), institutions on the whole were reluctant to abandon their traditional reliance on SC results, and sought additional, not alternative, ways to admit students.

An option that seems not to have been seriously considered at this stage was that of increasing the numbers of black students through random selection: that is, by establishing a lottery system to 'select' black students who did not meet the automatic point score requirements. Indeed, the overriding concern of controlling, in some way, the level of risk faced by the institution in admitting academically under-prepared students, would have made this option highly unattractive. In addition, UCT, which makes no bones (now or then) about viewing itself as a highly selective institution, would not have been predisposed to take seriously a non-merit based solution to the problems of selection.

In response to these concerns, the institution's Academic Planning Committee resolved in 1986 to:

- Ensure clear identification of the target group, in order to avoid the problems described by Cloete and Pillay. The group identified was students writing under the authority of the DET[51], as this group was the most educationally disadvantaged. In 1991, for example, for every R1.00 spent per pupil by the DET, R4.60 was spent by the white departments, R3.20 by the House of

[51] The target group was identified by educational background rather than 'race'. Any student writing the SC under the DET was eligible for the scheme, whereas black students who were not in the DET system were not eligible.

Delegates (the Indian department) and R2.60 by the House of Representatives (the Coloured department) (Bot 1994). While the relationship between resources and quality is complex, the real, cumulative effects of severe under-resourcing experienced by the DET system in particular, cannot be denied. In addition, all DET candidates are speakers of English as an additional language – and have had English as a medium of instruction through most years of schooling. As was argued in Chapter Six, difficulties with the medium of instruction play an important role in the creation of educational disadvantage.

- Create, by some means such as additional tests (additional to the SC examination), a ranking of results that would allow for discrimination amongst scores. At this stage it was believed that a large part of the problem of the restricted range of scores obtained by candidates on the DET SC would be overcome by a more rigorous assessment system. However, it rapidly became clear that a new approach to assessment was needed, rather than simply improved testing procedures. As Yeld and Haeck (1997:9) point out, traditional testing approaches tend to elicit " ... fairly uniformly dismal performances ... [which] ... blur the distinction between better and weaker candidates".
- Use, as the basis for these additional tests, skill areas believed to be of general value in the curriculum. 'Language' was agreed in the original design of the project to represent a crucial and core area of academic ability, and thus it was decided that the early efforts would entail the development of tests in this area. However, it was clear that the tests were a means to an end. That is to say, the project's decision to develop tests in these areas was opportunistic rather than principled: there existed no clear articulation of concern about students' abilities in these areas that would have resulted in their being tested had the SC been shown to be reliable and useful. This point is important, as it explains why the early test development and validation concerns were almost exclusively focused on the predictive, and not the construct, validity of the tests.

While it has not been clearly documented, the rationale for testing language arose in the following context. First, its inclusion in many testing initiatives elsewhere (e.g. the SAT in the United States, and various well-known language tests such as the TOEFL, the English Language Testing Service

(ELTS), as the IELTS was then known, and the Test in English for Educational Purposes (TEEP) in the UK, all of which were widely used for admission to overseas institutions), meant that there was already established practice on which to draw. This precedent was reinforced by the general perception of language (medium of instruction) as a core area of knowledge and skill, which made it a logical choice on which to base testing. In addition, the fact that English, the language of learning at UCT, was not the first language of the vast majority, if not of all, educationally disadvantaged DET students suggested that its inclusion in a testing project might be seen as legitimate. Finally, a tradition of language proficiency testing already existed at the institution. From 1984, all first-time entering students in the Faculties of Architecture, Arts, Commerce, Science and Social Science who had, in the SC examination, either written English as a Second Language, or as a First Language and obtained less than 50%, were required to write the English Language Proficiency Test (ELPT) developed by English for Academic Purposes staff in the then Academic Support Programme. Performance on this test was used to place students onto a credit-bearing English for Academic Purposes semester course. The existence of this test, which was never validated, set the scene for the development of the new AARP language tests in more ways than one, as is shown below.

Despite its ambitious aims, however, the scope of the project – in staffing and financial resources - was tiny. Staffing was minimal and temporary, consisting of the equivalent of two full-time staff (one full-time co-ordinator cum mathematics test developer, one one-third time language test developer [the author of this study], and one statistician on a two-thirds basis). Indeed, the original proposal for the project envisaged that once the first tests had been piloted, in 1987, the task would be to

> "... introduce the alternative procedure and policies in 1988 and carefully monitor the performance of students admitted in terms of them. Modifications might be made in 1989 and 1990 but by 1990 a cohort of graduates would be available for an overall assessment to be made" (University of Cape Town, undated:2).

This somewhat unrealistic assumption took no cognisance of the fact that educationally disadvantaged students, however talented, tend to take at least one year longer than the minimum time to complete a degree – and that in all likelihood, none of the students admitted through the

AARP in 1988 would have graduated by 1990. In addition, the need for assessment expertise was not recognised:

> "It is not envisaged that the appointees would be specialist educational psychologists or testers. The major requirement would be a broad understanding of tertiary education ... and intimate knowledge and experience of the local education system, particularly as it affects black students" (op cit:2).

This lack of recognition of the importance of technical expertise in the area of assessment reflected the mood of the times, and the urgent need for advocacy in relation to the admission of black students. In 1985, UCT had only 339 black students (3% of its student body of 11,844), and for many at the institution, this low figure was not seen as a problem. As is frequently the case, the executive and senior staff were some way ahead of their colleagues in this regard, and project staff found it necessary to expend considerable time and energy on advocacy work related to the aims of the project. As is clear with hindsight, however, the lack of technical expertise significantly impeded the test development initiative, and the staff appointed found themselves having to enskill themselves with some urgency – and considerable difficulty, given the prevailing climate.

Thus, much time and energy in the initial years went into establishing the legitimacy and logistical feasibility of the project, within and outside the institution. In 1987, piloting of the testing scheme was carried out in the Western Cape. Until this time, the only testing of incoming students that had taken place had occurred after they had been admitted. As Badsha et al (1987:5) comment: "... consultation is extremely difficult under a State of Emergency". Indeed, for much of 1987, township schools were virtually under a state of siege. Scholars were regularly assaulted or intimidated by security forces and vigilante groups, and spent a great deal of time organising their resistance to the state. Meetings with scholar representative groups were fraught and logistically difficult[52]. The effort of gaining acceptance from the student leadership and the community, however, was critical to the future success of the project. Understandably, given the climate of the times, black students were wary of subjecting themselves to further testing. Not only had their previous assessment experiences in the DET system been dismal and often corrupt, but the

[52] For example, project officers would be directed to a street corner in a township where barricades of burning tyres (erected as a deterrent against the police in their heavily armoured vehicles) made navigation exceedingly complex – on

knowledge that few of their peers had gained admission to UCT in the past led them to be somewhat cynical of the value of the exercise[53]. School principals, for example, expressed the opinion that this testing was merely another way that UCT had devised to exclude black students. Indeed, the restriction of the testing enterprise to DET students lent credibility to this view. In response to the suspicions and fears voiced by the community, the following undertakings were given:

- All testing would be on a voluntary basis. Any applicant who wished to be considered on the basis only of her/his SC results would not be penalised for this choice.
- Whereas good performance on the AARP tests would be taken into account in assessing a student's application, poor performance on the tests would not be reported to admissions officers. That is, project staff would only forward the names and results of candidates who had performed well on the AARP test/s to the admissions officers. This undertaking had the serendipitous effect of ensuring that AARP would have a less truncated sample for validation purposes, in that many students with poor test scores were admitted on the basis of their SC results. It was, however, resented by admissions officers who felt, with some justice, that project staff were knowingly concealing the names of highly at-risk students[54].

The table below shows the numbers of students who have been admitted to the institution and were recommended by the AARP Project over the years 1988 - 1999[55]. 660 of these students have graduated, and 104 have obtained a post-graduate degree or diploma.

Intake year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	Total
No. registered	20	55	142	137	189	214	188	200	195	275	159	144	1988

Table 7.2: AARP Registered Students

reaching the street corner, a new meeting place would be given, and eventually, the meeting might take place – or might not, depending on the ability of the student leaders to make the assignation.

[53] These views were expressed in meetings with the student leaders, and with school principals.

[54] In 1997 the undertaking not to reveal poor AARP test performance was removed, and all names of candidates, and their results, have been submitted to admissions officers since that time.

[55] As a rule of thumb, about one eighth of the students who are tested actually register at UCT. However, approximately one sixth of the students tested are made offers.

7.3 Test Rationale Issues

Previous chapters have stressed the importance of initial clarity about the future use of a new test, prior to any development activity. In general, such information must include, as a minimum, a clear understanding about the purposes of a test, about who will use its results, who the target candidates are, and what the resources and constraints are.

7.3.1 The Purposes of the Test

The original purpose of the PTEEP tests is outlined above; namely, to increase access for black, educationally disadvantaged students who are able to take advantage of the educational opportunities offered by the institution. Until the intake year of 1996, its statement of purpose read as follows:

> The PTEEP is developed by the Alternative Admissions Research Project to increase the numbers of black and/or educationally disadvantaged students at the University of Cape Town. The purpose of the PTEEP is to assess the potential of ex-DET applicants to the University of Cape Town to succeed in graduating from the institution. It attempts to do this by measuring the Cognitive Academic Language Proficiency of applicants, who are second-language speakers of English, in situations and tasks reflective of non-discipline specific university study at the institution. Candidates write the test on a voluntary basis, and will not have attempted previous tertiary level study. Abilities are reported as quintiles within rankings derived on the basis of pre-1996 educational arrangements. The quintile scores are intended to be used as one criterion in decision making for undergraduate admissions, in addition to other criteria such as Senior Certificate results. In addition, information derived from the scores may be used to assist in the making of placement decisions, and for the awarding of 'Senate's Discretionary Conditional Exemption' regulations[56] (Unpublished AARP document, 1994).

In more recent years, since the widening of eligibility to include non ex-DET candidates (described below in section 7.3.3), the following additional uses can be identified:

- to place students onto programmes of study, e.g. onto regular, extended or bridging programmes;
- for base-line testing of entry level skills, which could be used for course and curriculum design; and

[56] These regulations widened access to Higher Education as follows: " The Committee of Principals shall issue a certificate of conditional exemption to a person who, in the opinion of the Senate of a university, had demonstrated, in a selection process approved by that senate, that he/she is suitable for admission to bachelor's degree studies, which certificate shall be valid for admission to that university only" (amendment to Regulation 28 of The Matriculation Board's exemption requirements).

- to assist in the allocation of early offers of admission – that is, offers that are made before the institution receives the SC results. This is important in a context in which the internal school results – that is, results obtained from continuous and/or school-based assessments - submitted by applicants are not considered to be reliable, and where there is competition between institutions for good applicants;
- to assist in the allocation of scarce resources in the form of financial aid and/or housing places;

The current statement of purpose, in addition to omitting any references to the now defunct DET system, reflects the fact that the tests are no longer targeted only at second language speakers of English.

Nevertheless, the most important imperative of AARP was, and is, to provide an additional means of entry for talented candidates whose school-leaving results do not reveal their full potential, and who would not gain entry on the basis of their school results.

7.3.2 Test Users

A further area where clarity is needed concerns the identification of the people and offices that will use the test. In the case of a university admissions test such as the PTEEP, the following users and user groups need to be considered:

- Institutional users. These include admissions officers; financial aid officers, who need to allocate merit-based financial aid packages; lecturing staff who will teach the admitted applicants; and those in charge of quality assurance at the institution.
- Groups external to the institution. These include national DoE officials (who might wish to know about entry level skills across the system, for example); donor agencies who give bursaries, who might want to know that these are being distributed to worthy recipients, and the schools from which students are recruited.

Clearly, these user groups require different kinds of information. Admissions and Financial Aid officers need to know about relative merit, and about cut-points on ranks. Ideally, they also need

to know how the test information compares in terms of selection utility with other tests, such as the SC examination. Because selection was the overriding purpose of the new tests developed by the Project, the establishment of predictive validity was an urgent and paramount concern. Until admissions officers and Deans were confident that students selected on the basis of their performance on the AARP tests had at least a reasonable chance of success, they were understandably reluctant to make offers on this basis. Ironically, however, many admissions officers were prepared to make admissions offers to DET students on the basis of information they knew to be unreliable – i.e. SC results in the D or E aggregate ranges. Others were reluctant to make offers to any DET students at all unless they had satisfactorily high SC results, even though they knew that very few DET students obtained such results. As a result, in some faculties very few DET students were admitted, while in others the failure rates of such students were unacceptably high. In this context, the pressures were intense for AARP to establish a track record of success by recommending the admission of students who would be successful at the institution. This demand meant that the predictive validity of the tests had to be established as a matter of priority.

Many test users need to know more than simply a rank, however. If test results are to be at all useful to lecturing staff, for example, they need to know about the levels and kinds of skill and knowledge brought in by entering students, so that they can design their courses appropriately and so that successful applicants can be appropriately placed in curricula that meet their needs. This purpose raises a particular issue in relation to the development of the tests: that is, the level of performance at which detailed information about performance is thought necessary. For the PTEEP, since recommendations would not include candidates attaining scores below the top quintile, the tests need not be designed to deliver detailed information about the performance of candidates in the bottom categories. This type of summary outcome means that the tests would be designed with the majority of their items intended to discriminate amongst candidates in the top bracket of scores. As is discussed below, the widening of eligibility after 1995 to include all applicants to UCT, and not only ex-DET applicants, made this task far more difficult.

The different needs of user groups in relation to a test such as the PTEEP points to confusion about the purpose of the test. While the test developers were tasked with establishing a means of selection, and did so by producing a rank which clearly differentiated between prospective successful candidates and others, they soon came under pressure from colleagues in the Academic Support Programme, who wanted information about incoming levels of knowledge and skills for the purposes of curriculum and course design. Such tensions are still evident today.

The needs of those involved in quality assurance are similar to those of lecturing staff, except that the quality assurers might require the test to be structured in such a way that it can deliver information that is directly comparable with similar tests given at exit level (i.e. on graduation from the institution). In the early years of the Project, this need was not considered a priority, and so the tests did not include this design feature.

7.3.3 Test Takers

A third area that is essential for test developers to be clear about concerns the characteristics of the test-taker group. Issues to be resolved here include such matters as eligibility. Until 1996, the tests were restricted to DET students who had had no prior tertiary study experience. This eligibility was decided in an attempt to ensure comparability and fairness.

From December 1996, however, all South African schoolchildren attending public schools on a full-time basis have written SC examinations set on a provincial, not departmental, basis[57]. This move to a unitary system, organised on a provincial basis, while long overdue and widely welcomed, had many ramifications. Perhaps the most striking was the impact of writing a common examination on ex-DET and ex-DEC (Coloured) schools and students. The irony was that while the education, with some notable exceptions in both cases, offered to students in both these systems was markedly inferior to that experienced by students in historically white and Indian schools, the very segregation that made such discrimination possible also enabled gross inflation of the SC results

[57] School-leavers at public schools in the Western Cape had previously written one of four SC examinations, depending on the schools they attended. The four SC examinations were those of the Department of Education and Training (DET) for historically African schools, the Cape Education Department (House of Assembly) for historically white schools, the Department of Education and Culture (House of Representatives) for historically coloured schools, and the Department

within those systems, and protected African and Coloured scholars from direct competition with their more privileged peers. In order to lessen the blow of the common examination, the South African Certification Council - which is responsible for ensuring consistency of standards over various examining bodies from year to year – devised a 'desired distribution' of marks, which took the previous DET norms at face value. This shift resulted in a far higher number of A-symbols (marks over 80%) at previously advantaged schools, and added to public disquiet about the examination.

From the point of selection, the even more dismal performance of ex-DET students meant that far from the need for an additional means of access for ex-DET students becoming unnecessary, it became even more crucial, as the number of ex-DET students passing their SC in the new system was significantly lower than previously. At the end of 1997, for example, while the number of candidates who wrote the SC increased by 7% over the previous year, the number of candidates who passed decreased by 6%. This dramatic decrease can be seen also in the decline in numbers of students obtaining matriculation exemptions (Shindler 1998:1). The highest pass-rates occurred in the provinces with the lowest proportions of ex-DET learners, demonstrating the negative impact that the new system was having on ex-DET learners.

Eligibility for the scheme became more complex as a result of this change, however. It was decided to open eligibility to all applicants to UCT, irrespective of educational or language background, who had not had any form of tertiary education. The growing, and justified, resentment of coloured applicants (forcibly expressed by the principals of Coloured schools in several meetings with the Vice-Chancellor) about their exclusion from the admissions scheme was a further reason for the widening of eligibility. Eligibility remained restricted to South African students. However, this decision was taken on logistical, not principled, grounds, and any non-South African but otherwise eligible applicants to UCT who could appear at a testing centre were permitted to write. At this stage, too, the scheme remained restricted to students who had actually applied to UCT for admission.

---

of Education and Culture (House of Delegates) for historically Indian schools. In 1996, all scholars in public schools in the Western Cape wrote one examination, set by the Western Cape Education Department.

7.3.4 Resources and Constraints

Resources and constraints form a further area about which it is important to gain clarity prior to test development. In South Africa, with high levels of unemployment and poverty, it has been considered to be counter-productive to charge fees to write the tests, and so they are free for candidates. Clearly, this free service has implications for the affordability of the project, and has led to real tensions in test design. An example of a constraint in terms of human resources can be seen in the scarcity of available highly skilled people who could be employed as scorers, which in turn constrains the kinds of marking that can be undertaken. A further constraint is that of test security. Internationally, test security is a major concern, and is pronounced in South Africa with its high incidence of crime. For example, there is the constant fear of vehicle hijacking, which has occurred on two occasions to vehicles transporting project tests[58]. A more fundamental constraint can be seen in the design of the tests, which are all of the paper-and-pencil type. This form was chosen for logistical reasons: for example, not all testing centres have a reliable electricity supply, and so the tests cannot make use of sophisticated technologies such as computers or video recordings.

7.4 Conclusion

As is demonstrated in the following chapters, the development of the PTEEP tests faced certain distinct and difficult challenges. First, they had to elicit a range of performances from students whose educational backgrounds had the effect of depressing and, more importantly in this regard, homogenising test performances (Yeld & Haeck 1997, Haeck et al 1997). Drawing on the insights explored in previous chapters of this study, the approach adopted was that of task scaffolding. Its effectiveness is investigated in the following chapters.

[58] In both cases, fortunately, the tests had not actually been written. However, the tests were discarded by being thrown onto the side of the road, and left there – a real security nightmare.

Second, the tests had to address the issue of extreme diversity in educational preparation. That is, in one test, ranges of performance had to be elicited from two quite distinct (in terms of test performance) groups of candidates. This proved difficult to address in that, for example, items that 'worked' for one group tended not to discriminate at all for the other. The main strategy adopted in this regard was the separation of the two most distinct groups into discrete ranks, as is explained in Chapter Two (section 2.5.3). As the focus of this study, however, is on the performance of educationally disadvantaged groups, the discussion that follows concentrates on the validity of the PTEEP tests only in relation to the performance of ex-DET students.

CHAPTER EIGHT

CONSTRUCT VALIDITY
AND
THE PLACEMENT TEST IN ENGLISH FOR EDUCATIONAL PURPOSES (PTEEP)

8.1 Introduction

Cronbach (1988:3) suggests that: "[V]alidation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today, it is a public spectacle combining the attractions of chess and mud wrestling". In the United States, for example, the wide interest generated by the publication of books such as Lemann's *The Big Test* (Lemann 1999b), and the prevalence of litigation in respect of test scores indicate the increased importance of transparency and accountability. It is thus important for rigorous validation to be conducted, not only so that candidates and the public can be confident that fair judgements are being made, but also so that test developers and users can defend their tests and the decisions made on the basis of performance on these tests[59]. As Bachman and Palmer (1996:21) suggest, the aim of validation

[59] In many contexts, this justification is made on the grounds only of reliability. The Senior Certificate examination in South Africa, for example, lays considerable emphasis on the processes surrounding reliability, but its systems of moderation at all stages of the examination process are weak (Department of Education, 1998).

studies is to ensure that "... as test developers and test users we must be able to provide adequate justification for any interpretation we make of a given test score".

Messick (1989) points out that theoretical conceptions of validity have been gradually evolving from a focus on many different kinds of validity to a unitary validity conception. Different kinds (or types) of validity, as Alderson et al (1995:171) suggest, are viewed as "...different 'methods' of assessing validity", or as Cronbach (1988:4) suggests, as "... strands within a cable of the validity argument".

In the discussion of validity in the following chapters, these 'different methods of assessing validity' are categorised into three groups, viz. construct validity (this chapter), internal validity, i.e. content and face validity in Chapter Nine, and external validity (predictive, concurrent and consequential) in Chapters Ten and Eleven.

The essential requirements for a test of academic literacy that would be appropriate for admission to Higher Education in a context such as South Africa were outlined in Chapter Six (section 6.4.3). A prior requirement, however, is the need for clarity on the kind of test that is needed, and of the purpose or purposes for testing. In Chapter Seven, reasons for the establishment of the AARP Project were explored. Primarily, these were (and are) to devise selection criteria that would enable UCT to identify talented yet educationally disadvantaged black students whose SC results would not necessarily reveal their ability. Insofar as the development of the PTEEP tests is concerned (described in Chapter Nine, section 9.2), the purpose was not so much to assess the level of a candidate's language proficiency - for example to assess the candidate as 'beginner', 'intermediate', 'advanced' or as at a particular level in writing proficiency - as much as it was to predict performance in a future setting in which language is one of many variables. As such, it falls into the category of tests known as strong, or work-sample, tests (McNamara 1996).

Work-sample tests, as their name implies, derive their *raison de etré* from the work context in which successful candidates will be placed or are expected to perform. As has been discussed at some length in Chapters Five and Six in particular, the language-related work contexts in which successful candidates are expected to perform in formal education have several distinct

characteristics, which have become known collectively as academic literacy. Thus, tests of academic literacy in the higher education context must be based on a coherent understanding of learning and knowing relevant to higher education, and of the role of language within this. The model, articulated in Chapter Six, contains the underlying assumptions that, it is argued, should guide the development of the theoretical basis (the construct) of academic literacy tests in the South African context.

However, given the relative lack of certainty about what it is that should be tested, illustrated in preceding discussions on different conceptions of effective knowing and doing in Higher Education, it seems inevitable that tests of academic literacy will tend to conflate construct and content validity. As Weir (1988:17) suggests, in the context of the development of such tests, "[B]ecause we lack an adequate theory of language in use, *a priori* attempts to determine the construct validity of proficiency tests involve us in matters which relate more evidently to content validity." The conflation of construct and content validity is to some extent evident here in the assignment of certain validity investigations to Chapter Nine, which focuses on Content validity, rather than to this chapter, which deals with Construct validity. The principle underlying the allocation is this: the validity of those features of the tests that relate most directly to the testing approach adopted (viz. the inclusion of scaffolding devices, the insistence on active involvement, and the acknowledgement of the importance of content knowledge – the 'how' of the test) is investigated in this chapter, whereas the validity of those features of the tests that relate most directly to the ways in which the items have operationalised the language knowledge specifications – the 'what' of the tests - is investigated in Chapter Nine. Nevertheless, a degree of overlap is acknowledged.

Section 8.2 below illustrates the ways in which the PTEEP blueprint (framework and construct) for test design relates to the theoretically derived construct developed in Chapter Six. As a convenient shorthand, this latter construct is referred to below as the Prototype Construct, or PC.

8.2 The PTEEP Framework and Construct

In this section, an account is given of both the framework guiding the development of the PTEEP tests (viz. the aims, participants, and scope of the tests) and the construct (viz. the testing approach, the understandings of knowing and learning, and the notion of academic language proficiency or academic literacy on which the tests are based).

8.2.1 The PTEEP Framework

The main aim of the PTEEP tests is to predict the performance of candidates in future situations in which language will be an important, but not sole, variable. The tests set out to assess the academic language proficiency of all applicants, irrespective of study area. They are thus not discipline-specific. Their claim to be valid in these respects is investigated in Chapter Ten (Predictive Validity).

The PTEEP blueprint stipulates that the tests will only assess reading and writing, not listening and speaking. This is justified on the grounds of considerations of cost and capacity[60], and because reading and writing are so fundamental to the heuristic uses of language that assume such importance in tertiary study. Related to this, the framework stipulates that the tests will not assume the existence of ideal testing conditions (for example, it will not be assumed that electricity will be available): the implication of this is that the tests will be of the pencil-and–paper variety, able to be written with minimal supervision, in almost any conditions. These requirements have serious consequences for test design, as will be seen in the following chapters.

The PTEEP construct draws no distinction between first and second (or additional) language speakers of English. The complexity and reality of multilingualism in South Africa, it is argued, and the complex interaction of this with inequalities in educational provision, make attempts to 'classify' students on the basis of language background alone extremely difficult to implement or justify. In the case of the PTEEP (as was described in Chapter Two, section 2.5.3), the distinctions that are

[60] The IELTS test, which does test listening and speaking in addition to reading and writing, currently charges £72 (R 846) per candidate – a sum of money well beyond the reach of most South African applicants. The costs of the PTEEP tests are born by the institution, not the candidates.

made in terms of interpreting score results are on the basis of educational, not language, background.

8.2.2 The PTEEP Construct

Because of the prevalence and severity of educational disadvantage in South Africa, and the adverse impact this has on the ability of candidates to demonstrate their underlying abilities, both the PC and the PTEEP constructs stress the need to 'acknowledge the effects on cognitive functioning of the quantity and quality of learning opportunities experienced by an individual', and to develop ways in which dynamic testing approaches can be incorporated. The PTEEP tests attempt to address this requirement by adopting an approach known as scaffolded instruction. The effectiveness of this attempt (their claim to validity in this regard) is investigated in 8.5.2 below. The PC states that appropriate tests will be 'based on a notion of knowing and learning which views learners as actively involved ... in creating and negotiating meaning in a wide variety of settings'. The extent to which the PTEEP tests can be said to have successfully created conditions conducive to 'active involvement' is investigated in section 8.3 below.

Conceptual development, according to the PC, is 'seen as highly dependent on specific areas of expertise involving knowledge and information, and on the connections between these'. The importance of topical knowledge is thus recognised as an important component of the PC definition, although it is emphasised that, as far as possible, this knowledge will be limited to information provided in the test (i.e. prior knowledge of the topic is not included in the construct). In order to provide real opportunities for cognitive engagement and for the demonstration of this, the PTEEP tests claim to be based on complex themes (thereby providing topical information), and to include demanding tasks (thereby providing opportunities for cognitive engagement).

In acknowledgement of the serious difficulties in assessing metacognitive strategy use (strategic competence) in a pencil-and-paper test, the PC states that this will not be directly assessed. Similarly, the PTEEP tests take seriously the PC undertaking to acknowledge the impact of affective schemata on performance, and claim to make efforts in the design and layout of the tests, as well as in their administration, to minimise negative effects as well as to promote and harness

positive effects. These claims are investigated in Chapter Ten (Content, Face and Response Validity).

The PC is based on a componential notion of language ability, closely modelled on that of Bachman and Palmer (1996). This comprises topical knowledge (as mentioned above) and language knowledge, mediated by strategic competence, or metacognitive strategy use, and affective schemata. In Figure 8.1 below, the middle and left-hand columns contain a representation of the Bachman and Palmer (1996) model of language knowledge, which was displayed in Chapter Six (Figure 6.2). The column on the right contains the specifications developed for use by the PTEEP test developers. These represent the aspects of language knowledge that are believed to be particularly essential in an academic context, and which are feasible to test in the PTEEP context. Thus, not all of the aspects listed in the Bachman and Palmer model are included in the PTEEP model. For example, morphology (listed under Bachman and Palmer's 'grammatical' category) is not included as a PTEEP specification.

In the development of the PTEEP, two 'new' kinds of ability were incorporated into the model of language knowledge proposed by Bachman and Palmer. The first, implied but not explicitly mentioned in the Bachman and Palmer approach, concerns the kind of ability required to decode or construct graphs, tables, diagrams, maps, flow-charts, and other kinds of non-linear information presentation. For many students, difficulties with these ways of representing information mean that they cannot fully access the information in a text. Tables and graphs are simply ignored, or sometimes misinterpreted, often with serious consequences. This type of ability is included in the table below as an enabling skill in the category of functional knowledge.

The second kind of ability does not commonly appear in taxonomies of language knowledge. It, too, however, represents a crucial source of knowledge contained in text (verbal or written). It involves the kind of ability required to understand and use numerical references such as those commonly encountered in academic contexts. These references include understanding basic numerical concepts and/or information used in text, plus basic numerical manipulations such as estimation, comparisons (e.g. greater than, smaller than), percentages, basic fractions, basic

chronological references, sequencing, and basic computations. For example, in a Sociology lecture, students need to be able to understand references such as "Five years ago, twice as many people were living in conditions of poverty as the 30% in that category now". Examples abound in such disciplines as History (where chronology is important), and Social Work (where such matters as trends, or budgets, have to be considered), as well as in the more obviously numerate discipline areas such as Economics, Biology and Chemistry. This type of knowledge, 'knowledge of numerical concepts in text', is included in Figure 8.1 below as a subset of functional knowledge.

The use of specifications in the development of tests is somewhat controversial. It can lead, for example, to an overly mechanistic, 'check-list' approach, which can detract from the authenticity and natural logic of texts and tasks (Barsby et al 1994). Nevertheless, the existence of specifications, if used judiciously, provides both a framework and a safeguard against an overemphasis on certain kinds of task or skill. From the perspective of the PTEEP construct, the specifications contained in the right-hand column of the table below represent the essential elements insofar as the construct of language knowledge is concerned. The extent to which the PTEEP tests can be considered to be valid in respect of the model of language ability on which they claim to be based is investigated in Chapter Ten (Content Validity).

Language Knowledge (Bachman & Palmer 1996)		PTEEP Language Knowledge Specifications (Yeld et al 1997)
ORGANISATIONAL KNOWLEDGE	Grammatical Vocabulary Morphology Syntax	Vocabulary: 'unknown' vocabulary (deriving meanings from context); 'known' vocabulary (i.e. no context provided); spelling as it affects meaning Syntax: understanding the syntactical basis of the language
	Textual Cohesion Rhetorical organisation	Understanding relations between parts of text (e.g. through devices of cohesion such as pronoun reference, particularly demonstratives, referring to statements/propositions or 'entities', and/or by recognising indicators in discourse, especially for introducing, developing, transition and conclusion of ideas, and signalling relations between phenomena). Skimming and scanning (e.g. using macro features of text such as headings, illustrations) to get gist of passage, locating particular pieces of information Extrapolation and application (e.g. drawing conclusions/applying insights derived from texts, seeing trends) Inferencing: (understanding ideas/information in a text, implied but not explicitly stated).
PRAGMATIC KNOWLEDGE	Functional knowledge Ideational Manipulative Heuristic Imaginative	Separating the essential from the non–essential (e.g. main idea from supporting detail, statement from example, fact from opinion, proposition from its argument, classifying and categorising). Detailed reading for meaning, at sentence level and at discourse level Understanding the communicative function of sentences with or without explicit indicators, such as definition, exemplification, exhortation, argument/persuasion Understanding the importance of 'own voice' (including 'ownership' of ideas) and/or creativity of thought and expression Knowledge of visually encoded forms of information representation (graphs, tables, diagrams, maps, flow-charts) Understanding basic numerical concepts expressed in text/numerical manipulations (comparisons, e.g. greater than, smaller than, percentages, basic fractions (e.g. half of, more than double), basic chronological references, sequencing, basic computations
	Sociolinguistic knowledge (sensitivity to dialect, language variety; register; naturalness criteria); familiarity with cultural references and figures of speech	Understanding metaphorical expression Understanding text genre (including audience, purpose etc.)

Figure 8.1: PTEEP Language Knowledge Specifications

According to the PC, the tests must be based on a sound understanding of academic tasks, which will require candidates to demonstrate a wide range of relevant abilities (Bachman and Palmer's notion of 'ability-tasks' is relevant here. This is elaborated in the PTEEP construct as meaning that the tasks will require the ability to comprehend information presented in various modes, to paraphrase, to present information visually, to summarise, to describe (e.g. ideas, phenomena, processes, changes of state), to write expository prose (e.g. argument, comparison and contrast, classification, categorisation), to develop and signal own voice, to acknowledge sources, and to perform basic numerical manipulations. In demonstrating these abilities, candidates will be required to construct and write summaries, write expository prose in the form of a one-page essay in which they adopt and support a position, drawing on the information provided in the texts, to construct and read graphs, flow-charts, and diagrams, and perform simple numerical manipulations within the context of the test's theme. They will in addition be required to answer a number of questions assessing reading comprehension.

Finally, both the PC and PTEEP blueprints state that the tests should use a variety of item formats. This finds form in the PTEEP in a mixture of multiple-choice, constructed response and extended writing formats, as is demonstrated in Chapter Ten.

Sections 8.3 – 8.5 below illustrate the ways in which the PTEEP testing approach attempts to incorporate the insights on learning and knowing expressed above. Both section 8.3, which deals with the extent to which the test incorporates opportunities for active involvement of the candidates, and section 8.4, where the role of content knowledge is examined, rely on examples and illustrations from the test in order to assess the extent to which the test can be considered valid in these respects. Section 8.5, on the other hand, uses both quantitative and qualitative techniques in investigating claims of validity in respect of the scaffolding approach.

8.3 Active Involvement of Candidates

An essential requirement articulated in the construct is that candidates should be actively involved. In a sense, of course, test takers have no real choice about becoming actively involved if they are

at all serious about their performance on a test. However, the design of tests – the item formats they contain, the sequencing of items, the 'user-friendliness' of the item rubrics, the texts themselves - can impact dramatically on how an individual engages with the test. For example, a test that relies entirely on multiple-choice items shapes candidates' involvement in ways that are different to those of tests that include a mix of constructed response and multiple-choice items[61].

Whatever the particular format chosen, test questions draw on different combinations of what Weir (1983) terms 'enabling' skills. This is illustrated below, where the example of the task of summarisation[62] is used to show how the test questions elicit different enabling skills and draw on different cognitive processes involved in the act of summarisation. In an attempt to tap as many of these processes as possible, each PTEEP test includes a variety of summarising tasks as shown below:

- complete a 'summary cloze': Cloze is a technique in which words are deleted from a text, and are supplied by the candidates. In language tests, the deletion is usually on an n^{th} word basis, or on the basis of a particular grammatical form (e.g. all prepositions could be left out). In the case of summary cloze, the omitted words have been deliberately selected to represent key points of a text. The primary cognitive activities the candidate is engaged in for this task are those of identification and selection of the required pieces of information. In line with the aim of the tests to provide as much contextual support as possible, however, it needs to be noted that in completing the task, candidates are in fact constructing a summary, with considerable support, that will be of use later in the test – for example, in writing the final essay. This scaffolding strategy occurs in one form or another after each piece of text in the tests.
- summarise a specified text or text extract (e.g. three paragraphs setting out a particular argument) in a set number of words. In this case the candidate must use all three of the broad

[61] As Skakun and Maguire (2000) point out, multiple-choice items require test takers to make selections from a number of alternatives.. These alternatives can be used in a hypothetico-deductive fashion, where candidates generate hypotheses and accept or reject alternatives; as a memory cue which can act as a support for the candidate; and as a focusing agent which alerts the candidate to possible forms of response. These are rather different forms of engagement from thase required by constructed response questions, and point to the need for a mix of formats

[62] The different cognitive processes involved in summarisation were briefly noted in Chapter Six. These processes are selection, condensation and production (incorporating integration, combination and transformation).

cognitive operations involved in summarisation suggested by Hidi and Anderson (1986), namely selection, condensation and production.

- construct or complete a flow-chart which summarises the main points, or framework, of a text. In addition to the three processes referred to above, this task requires candidates to transpose information from one form (prose text) to another (diagrammatic). This requirement places a greater emphasis on the transformation of the material.
- extract the main points of a particular topic or argument in order to achieve a particular communicative purpose (e.g. a letter to an editor supporting a development project). This task is somewhat similar to the transposing summary task described above, with a major difference - a sociolinguistic dimension is added through the inclusion of an audience for whom the summary must be tailored.

The range of summary tasks described above illustrates the emphasis given to creating many and varied opportunities for candidates to engage actively with the texts and tasks within the tests, in accordance with the importance in the PTEEP construct of active involvement of candidates. A similar range of tasks is included to tap other skill areas or groups of enabling skills, as is demonstrated further in Chapter Nine, sections 9.2 and 9.3. While validity in this aspect is not quantified, it is argued that the range of tasks, activities, and types of engagement required of candidates supports claims of validity in this regard.

8.4 Content Knowledge

Another important way in which the PTEEP attempts to incorporate the understandings of knowing and learning expressed above in the model can be seen in the recognition of the importance of content (or topical) knowledge and information. This incorporation is achieved through the use of a theme. A variety of texts (e.g. academic articles, illustrations, tables and charts, brochures) is used, based on a topic which is chosen so that it is new to candidates, or treated in a novel way, yet will provide as much complex information and contextual support as possible. All of the PTEEP tests have the following content characteristics in common. The tests begin with a 'fact file' – a structurally simple, short text that lays a factual foundation. Graphs and tables illustrating and

building on the information in the fact file follow this. Next, an issue relevant to the theme, but situating it in a South African context, is provided[63]. The 'issue' texts selected are written in an academic style, and are fairly complex. The last text in each test is a piece of popular – that is, less academic - writing.

The provision of a theme that is both conceptually rich and varied in terms of genre assists, it is argued, in creating opportunities for candidates to 'create and negotiate meaning in a wide variety of settings'. The ways in which the PTEEP tests incorporate content knowledge and information, it is therefore argued, makes it legitimate for validity in this respect to be claimed.

8.5 Incorporation of Dynamic Assessment Principles

Taking seriously the impact - on cognitive functioning and test performance - of the quantity and quality of mediated learning opportunities experienced by an individual requires test designers to be cautious about making easy assumptions concerning a candidate's underlying ability from her/his test performance. In the context of widespread educational disadvantage as well as enormous diversity in educational preparedness, the PTEEP test development process is particularly focused on developing methods of eliciting optimal performances from all candidates, irrespective of educational background.

Earlier chapters have discussed some of the ways in which theories and practices in cognate areas have contributed to the development of assessment procedures that could complement traditional testing approaches. In particular, the principles of dynamic assessment were explored, and it was suggested, for example, that the Vygotskian notion of the Zone of Proximal Development (ZPD) could possibly be adapted to complement traditional approaches to the testing of academic literacy. Section 8.5.1 below describes how the PTEEP tests have attempted this.

[63] Experience has shown that unless there is a clear link to the South African context, test users are sometimes tempted to characterise the test as "Eurocentric" and dismiss its usefulness on those grounds. An example of this linking, from the 'Antarctic-PTEEP', is an expository piece of writing on the issue of whether South Africa can continue to spend scarce resources on maintaining its base in the Antarctic, while resources are so desperately needed in the country for basic needs such as housing.

8.5.1 The Scaffolding Approach

The approach adopted by the PTEEP in its attempt to provide opportunities for candidates to engage in activities that both encourage and reveal concept and skill development takes the form of what has been termed a `scaffolding' approach (Yeld & Haeck 1997, Palinscar 1986, Spring, Sassenrath & Ketellapper 1986). In this approach, exercises (opportunities for action) are created which lead candidates to manipulate the material (e.g. texts) on which the test tasks are based. The notion of scaffolded instruction in formal education is not new. As Palinscar (1986) points out in a review of relevant literature, it is what good teachers have always done, embodying as it does "... the best of teaching practices, by directing attention to learner and skill profiles, and to ... matching the skills of learners to the way new skills are presented ... "(Palinscar 1986: 95). Its use in a large-scale pencil and paper test is, however, novel. This `scaffolding' makes it possible for candidates to engage with the test tasks in ways that are different from those which would have been employed had the scaffolding exercises not been worked through (Yeld & Haeck 1997). The notion of scaffolded instruction is thus based upon the notion of the ZPD, as well as on Cummins's portrayal of the interaction of contextual support and cognitive demand. That is, by engaging with a task (a problem) under the guidance of task scaffolding, the candidate's engagement with the task will be enhanced, and her/his ability revealed more effectively than would otherwise be the case.

Various examples of scaffolding from the PTEEP tests are given below to illustrate the approach. At this point the examples and assumptions are simply described. Section 8.5.2 below contains an analysis of the ways in which the effectiveness of the scaffolding approach has been assessed.

Example One:

The first example is drawn from the 'Water-PTEEP' (i.e. the PTEEP test based on the theme of 'Water'), which was used in the 1999 entry cycle. The text contains a bar graph displaying how many litres of water were required, per kilogram, to produce various foodstuffs. For example, the graph reveals that it takes 24,640 litres of water for the production of one kilogram of beef, and 6,600 litres for one kilogram of chicken. Candidates then answer a number of questions that culminate in the selection of a 'waterwise' menu. The culminating question is shown below:

You have been asked to develop a menu for a meal, using foods that have taken as little water as possible to produce. Which of the following three menus would you select? First calculate the litres of water each menu requires, and then circle the menu that requires least.

Menu 1	Menu 2	Menu 3
250g chicken	100g beef	500g eggs
100g white rice	100g brown rice	200g brown rice
100g broccoli	200g tomatoes	800g melon
..... litres of water	..... litres of water	..... litres of water

As it is known that many candidates, and in particular educationally disadvantaged candidates, have great difficulty with numerical manipulations, the question above is first scaffolded as follows:

Question	Comment
1. How much water does it take to produce 1 kilogram (1,000g) of chicken?	This is a straightforward reading (transposing) task. The item acts as a warm-up opportunity, and confidence booster, as well as providing the basis for item 2. It also alerts candidates to the fact that 1 kilogram = 1,000g.
2. How much water would it take to produce 200g of chicken?	In this task candidates need to recognise that 200g is one fifth of 1,000g, and to take the number of litres they had read off the chart for item 1 (6,600) and divide it by 5.

After working through lead-up questions of this kind, candidates are, it is hypothesised, in a better position to deal with the waterwise menu question. They have had an opportunity to 'read-off' the graph, and also to perform a basic calculation using such information.

Example Two:

It was stated above (section 8.4) that the topical information in the test – that is, the textual content - is embodied in a theme that presents new yet sufficiently challenging information designed to provide a basis for complex cognitive processing. One of the devices used to 'scaffold' the texts is the use of a series of True/False questions. These questions, although the candidates are not informed of the fact, are not assigned any marks. Candidates are not informed that the T/F questions will not be marked so that they will take the questions seriously. The reasons for not marking them are (i) the item format lends itself to guessing and is generally avoided unless negative marking is employed, and (ii) as the questions are meant to scaffold understanding, rather than to test it, the great majority of the candidates get the items correct - this makes marking them inappropriate as one of the main aims of the tests is to discriminate between weaker and stronger

candidates. The purpose of the questions is to ensure that candidates engage with various core concepts or pieces of information in the texts. These concepts or pieces of information form the foundation for further texts and tasks in the test, and it is believed that inclusion of these items assists in providing a structured reading experience. Drawing once more on the 'Water-PTEEP' for illustrative purposes, the following True/False set shows both the straightforward yet core nature of the questions.

	T/F
Water shortages are found only in areas of drought.	
There is less water on earth today compared to millions of years ago.	
The demand for water has increased over the years.	
Industrialisation has not affected the demand for water.	
Drought conditions always cause disastrous famines.	
There is less water (per person) on earth today compared to millions of years ago.	
People have unequal access to water.	

Example Three:

In a similar fashion, the use of the 'summary cloze' (referred to above in section 8.3) creates a situation where the candidate needs to engage with the core concepts of the text or texts in order to complete the task. In this sense the surrounding text acts as a supporting context and an additional opportunity to engage with the text. The words omitted represent core concepts or information in the text, and cannot all be supplied without reference to it. An example of this is given in the cloze extract below, taken from the 1996 'Education-PTEEP'.

..... is believed, by many people, to be both the best, and the only possible, type of education. This is a ....., partly because there is simply not enough ..... to provide schooling for all, and partly because it is based on ....., which means that some pupils must fail, and others succeed. Education has thus been ..... by schooling, which has made it into a ..... with a university degree as the final prize, and lots of examinations along the way. (extract)

Example Four:

A further example of scaffolding can be seen in the ways that the extended writing task is developed. For example, an illustration of the task immediately prior to the essay In the 'Fire-PTEEP' is shown below. The purpose of this task is twofold: to help to familiarise candidates with the notions that each article puts forward distinct core concepts, and that sources should be acknowledged (the candidates are instructed in the essay rubric to do this); and to remind them of

some of the core concepts on which they will be basing their essay. A number of these items are given, however, not all of which are directly relevant to the essay, so candidates cannot draw on this task for the essay without applying their minds. The example below shows one of these items.

Read the following statements and identify, in each case, from which source (which article) the information comes.

South Africa has many urgent development needs which raise questions about its ability to afford its activities in the Antarctic.

Introducing Antarctica	
Assessing the Costs	
The Need for a Conservation Plan	
Explore Antarctica	

Example Five:

In some instances, scaffolding takes the form of direct instruction. In these cases the aim is to achieve optimum clarity about what is being expected rather than to provide modelling or preparatory activities. The example below is taken from the 'Radio-PTEEP'.

A metaphor is a figure of speech where one thing is described in terms of another. An example could be "he wept an ocean of tears". In this example, the word 'ocean' tells us that he cried a great deal, although of course his tears could not possibly really amount to an ocean's worth.

Which <u>two</u> of the following words from Paragraph 1 are used metaphorically? (Make two crosses on the Answer Sheet).

isolation	A	skeleton	C	handful	E	low-tech	G
technology	B	hammer	D	upgrade	F	nail	H

The decision to include such scaffolding in a test rests on the belief that providing appropriately targeted mediation with which candidates are 'forced' to engage will in turn change the way a candidate is able to engage with test tasks. The change which would occur as a consequence of engagement with scaffolding activities would, it was believed, address the two major selection problems identified at the close of Chapter Seven. The first of these problems concerns the generally very low level of performance of ex-DET candidates on the tests in use at this time. Admissions officers were understandably reluctant to make offers to candidates obtaining less than 40% on a test, whether or not they were at the top of their candidate pool. The second concerns

the clustering of the test scores of DET candidates, which makes selection extremely difficult and arbitrary. In other words, the intention of incorporating scaffolding was both to increase the level of performance, but, crucially, to spread the scores. It was therefore not the intention simply to make the tests easier for all candidates.

Four examples of the ways in which scaffolding has been incorporated into the PTEEP tests have been described in this section. The validity of the scaffolding approach employed is assessed in 8.5.2 below.

8.5.2 Assessing the Effectiveness of the Scaffolding Approach

In 1991, the new approach to testing - that is, incorporating the scaffolding approach - was introduced. At this time, the existing English Language Proficiency Test (ELPT) that had been used for several years to place students onto an English for Academic Purposes course in the Faculties of Arts and Social Science, needed to be replaced.

In order to make it possible for the effectiveness of the new approach to be investigated, the new test (the PTEEP) was deliberately designed to be identical in key respects to the test it replaced, the ELPT. Specifically, the two tests used the same texts and text types (prose, graph and diagram), and required the same three major writing tasks - summary, description and contrast/comparison. The difference between the two tests, in terms of their construction, was that the PTEEP included structured and sequenced tasks designed to act as mediation for the writing assignments.

The two tests were not, however, written by the same candidates. Given the difficulty in persuading students to write the tests once they had been admitted to the institution, this was not attempted. However, in Chapter One, section 1.4.3, it was argued that if certain conditions were met, it could be assumed that there are not significant cohort differences in respect of PTEEP testing. In the ELPT/PTEEP situation, these conditions are met. That is, the numbers were comparable and large, with 897 writing the ELPT and 1,153 the PTEEP; both tests were written in the final school year, immediately prior to the writing of the Senior Certificate; and both sets of

candidates were registered at DET schools when they wrote. In addition, the tests were written for the same purpose and thus it can be assumed that motivation in both cases was comparable.

The main questions that need to be addressed in assessing the effectiveness of the scaffolding approach are:

- Did the approach provide a greater *range of scores* (spread), so that capable students could be more clearly differentiated from weaker students?
- Did the approach increase the *predictive validity* of the test (that is, did the test correctly distinguish between weaker and stronger students)?
- Did the approach *improve* (raise) *the level of stronger students' scores*?

Three sources of evidence indicate that the scaffolding approach has made a difference to the way in which students perform on the test, and provide at least partial answers to these questions.

8.5.2.1 Range of Scores

The aim of this study was to investigate whether the introduction of scaffolding had increased the spread of scores. To assess this, the marks of the summary, descriptive paragraph, and essay were calculated, for both the ELPT and the PTEEP, as though they were the only items in the test in each case. These items are the same in both tests.

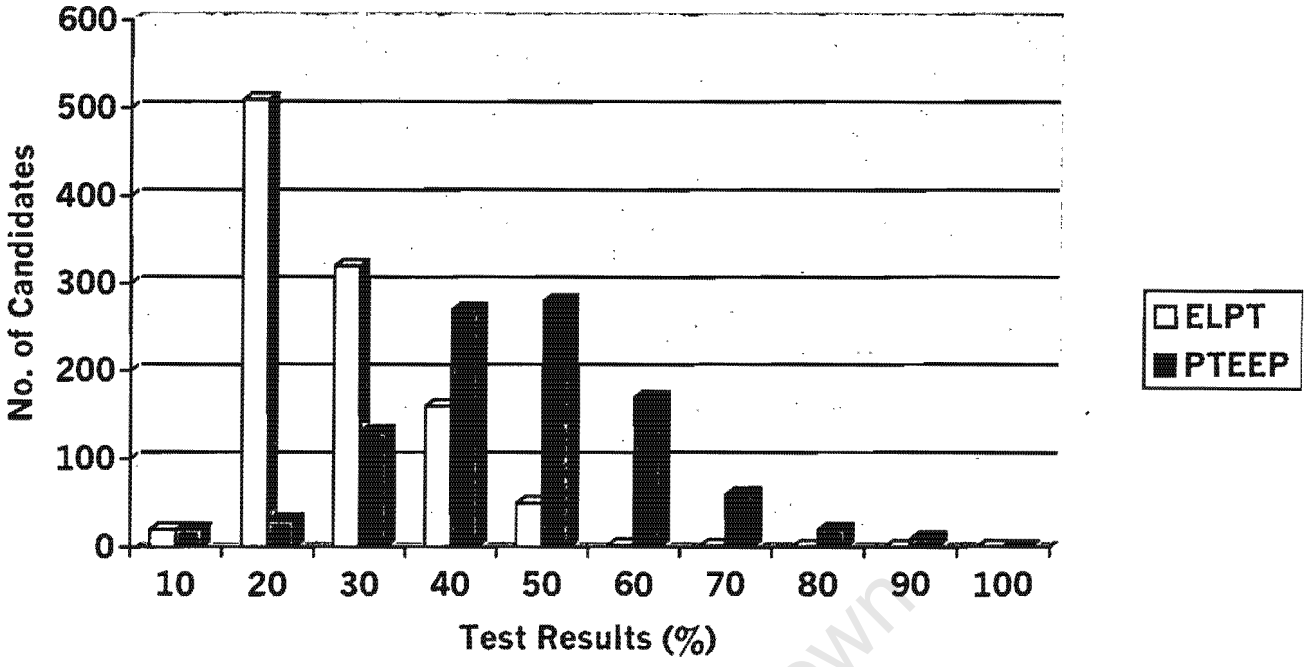

Figure 8.2: ELPT and PTEEP Scores

It can be seen from this figure that the new approach made a marked difference to the distribution of scores. The greater range of scores achieved by the PTEEP candidates indicated that the new testing approach was making a difference to performance. This increased spread addresses the major problem for selection of having what Yeld and Haeck (1997) have termed "uniformly dismal" performances from which to select. This upward shift in distribution on its own is an important finding, as the establishment of a significant right-hand tail means, at least in a technical sense, that selection can be undertaken on a less arbitrary basis.

Also, the test was not simply made easier: that is, the distribution did not simply move to the right, but reflected a wider range of scores. If it simply became easier (to an equal extent) for all candidates, all scores would improve by the same absolute amount - the curve would display the same narrow range of scores, but would shift to the right. This would not make selection less arbitrary. This feature suggested that the scaffolding worked differently for different candidates, in that weaker candidates did not benefit to the same extent as their stronger peers. Such a conclusion can only be supported by bivariate data on ELPT and PTEEP on a single set of candidates. Nevertheless, the data are suggestive, and encouraging in terms of supporting the value of the scaffolding strategy in increasing the range of scores.

8.5.2.2 Task Preparation

In assessing the impact of the scaffolding approach on the performance of individuals, Yeld and Haeck (1997) investigated the performance of students in the top and bottom three deciles (i.e. approximately the top and bottom 30% of the candidates) in overall PTEEP test score (N = 114) on the following questions:

> ***Question 6.1**: In note form (i.e. not full sentences), list the main points the author makes about the Japanese educational system. For this exercise, do not include comparisons or references to the American system.*
>
> ***Question 6.2**: Drawing on the text of the whole article, but using your own words, summarise the points the author makes about the Japanese system of education. Note: your summary should not include comparisons or references to the American system.*

The aim of the investigation was to assess the extent to which the note-making task (question 6.1 above) acted as preparation for the summary (6.2). The focus was thus on the number of points (correct or incorrect) listed in 6.1 and then carried over to 6.2. The study therefore examined (i) the number of points made in the notes and transferred to the summary, and (ii) the number of points made in the notes and not transferred to the summary. As might have been expected, the stronger candidates made more correct points in their notes than weaker candidates (r with overall test score is 0.39). In addition, they transferred more of these points to the summary ($r = 0.62$). However, weaker candidates, although making fewer points in the notes, transferred few of the points that they did make ($r = -0.35$) to the summary. The authors argue that weaker candidates did not connect the note-taking exercise with the summary, and suggest that the evidence points to the two tasks having been treated separately by these candidates, despite the very similar wording of the instructions. In addition, they note that better candidates on the test appear to have used the notes as direct preparation for the summary (Yeld & Haeck 1997:14).

Two tentative conclusions are drawn from this study. First, the authors suggest that scaffolding does not benefit all candidates in the same way, so the test is not simply easier for all candidates. The distribution of scores shown in Figure 9.1 above appears to confirm this claim. Second, they conclude that using this approach might in fact *dis*advantage weaker candidates, as it will widen the gap between them and their more capable peers. These findings are important, as a restricted

range of scores is a major problem for selection, and widening the gap between stronger and weaker candidates was an important aim of the new approach.

While this is not advanced in the Yeld and Haeck article, it could be argued that the findings also suggest that scaffolding of this type might have a role to play in the assessment of metacognitive strategy use, in that the use of the note-making exercise for the summary task requires both planning and judgement.

8.5.2.3 Predictive Validity and Scaffolding

From the point of view of selection, widening the gap between weaker and stronger candidates is a highly desirable outcome, provided, of course, that the test accurately distinguishes between stronger and weaker candidates in ways that correspond to performance in the criterion situation. In this connection, Yeld and Haeck (1997) report that correlations between first-year performance and the selection tests did in fact improve slightly with the introduction of the scaffolding approach, although this improvement was slight and the correlation, while significant, still very weak: from r = 0.23 (N = 372) in respect of the ELPT and performance, to r = 0.29 (N = 290) in the case of the PTEEP. The predictive validity of the PTEEP tests is discussed in detail in Chapter Ten, and so is not discussed further here, except to note that the evidence of improved predictive validity (of the PTEEP compared to the ELPT) provided some support for the new testing approach.

Clearly, much work needs to be done on refining and further researching the scaffolding approach. In particular, the implicit assumption that task scaffolding can impact on candidates' cognitive structures needs empirical investigation. Perhaps less ambitiously, but just as importantly, serious empirical work needs to be conducted to identify and develop the kinds of tasks that have most impact on candidates' performance in test situations.

Nevertheless, even in the absence of such research, the results point to an important role for task scaffolding as a strategy for testing educationally disadvantaged candidates. It appears to widen the range of scores, therefore allowing for more effective discrimination amongst candidates, it

appears to enable more effective discrimination between weaker and stronger candidates, and it appears to raise the level of stronger students' scores.

This has important implications, as previous attempts (e.g. those discussed in Chapter Four, section 4.3.2) to tap the underlying abilities of educationally disadvantaged students have relied on costly and time-consuming procedures. If further research confirms the usefulness of the approach discussed in this study, it will have major benefits for the identification of talented although disadvantaged students in a context where costly alternatives are not viable.

8.5 Conclusion

In this chapter, the construct validity of the PTEEP has been approached from many angles in an attempt to assess to what extent the tests succeed in measuring the notion of language proficiency that they purport to measure.

In summary, the analysis reveals that in terms of construct validity, the PTEEP tests are valid in many respects. The construct of the tests rests on several core features or assumptions. Several of these, such as the emphasis on ideational and heuristic uses of language, are addressed in Chapter Nine. In relation to the validity analysis in this chapter, the following features were investigated.

One of the most fundamental assumptions of the PTEEP construct is that the abilities of educationally disadvantaged candidates can most equitably and effectively be revealed or elicited by the adoption of dynamic testing approaches. The approach adopted in the PTEEP tests is that of scaffolding, and section 8.5.2 investigated the extent to which scaffolding can be considered to have achieved its aims of increasing the range of scores, and raising the level of scores for talented candidates. Drawing on data from various studies, it is demonstrated that the introduction of task scaffolding into the tests has made a significant difference to test performance: as outlined above, it has resulted in a greater range of scores, a generally raised level of performance for stronger candidates, and has strengthened the predictive validity of the tests.

The section concludes that, while further research is needed, the PTEEP tests can be considered to be 'dynamic' – that is, they can be regarded as valid in respect of this aspect of their construct.

Second, the construct holds that it is that it is through active engagement with a variety of tasks that candidates can most effectively demonstrate their ability. The discussion in section 8.2 draws on data and examples from the tests that illustrate the ways in which candidates are required to engage with the tasks, and the section concludes, on the basis of the data, that the PTEEP tests can be considered to be valid in this regard.

In addition, it is assumed that the most useful and valid representation of candidates' abilities in respect of particular skills or skill areas requires the provision of complex and novel topical information (embodied in a range of genres and text types) as well as a range of item types. The extent to which the PTEEP tests are valid in this respect is addressed in Chapter Nine, which deals with content validity. It is, however touched on in this chapter (sections 8.2 and 8.3 above), where it is argued that the provision of a wide range of text and item types can be regarded as meeting many of the validity demands of this construct feature.

Chapter Nine contains an analysis of content validity – that is, the extent to which the content of the test (the texts, tasks, rubrics, and marking scheme) constitutes a valid representation of the context.

CHAPTER NINE
CONTENT, FACE AND RESPONSE VALIDITY
AND
THE PLACEMENT TEST IN ENGLISH FOR EDUCATIONAL PURPOSES

9.1 Introduction

Various aspects of the construct validity of the PTEEP were analysed in Chapter Eight. In this chapter, three further aspects of the internal validity of the PTEEP are examined – viz. content, face and response validity. The construct of a test is largely defined through the specification of the knowledge, skills, and abilities that the test aims to measure. As was discussed in Chapter Eight, these specifications arise from an understanding of the "... complex of knowledge, skills, or other attributes ..." (Messick 1994:16) that are deemed to be important and likely to be encountered in the situations for which the test is considered appropriate. Test content, on the other hand, refers to the ways in which a test sets out to measure the specified knowledge, skills and abilities. Investigations into the content validity of a test are therefore investigations into the extent to which what is in a test (the texts, tasks, items) can be said to elicit the test's specifications. Test content, in other words, represents, or embodies, the operationalisation of the test construct. The challenge for content validity is thus to ensure that the behaviours and performances (the tasks and items) in the test, as well as the way they are scored, actually elicit and reflect this knowledge, skill, and other attributes. Face validity and response validity, in contrast, relate more directly to the ways in which a test is experienced or perceived by test takers

and users, and thus relies on different procedures and techniques for establishing validity. These two types of validity are discussed in sections 9.4 and 9.5 below.

However, the attempt to replicate, in tests, the kinds of tasks found in the criterion context raises the difficulty of generalising across different task types. For instance, instead of containing numerous multiple-choice items, a test might contain three or four tasks, such as writing an essay using multiple sources of information, or summarising an argument embedded in an extended piece of text. Linn et al (1991), in a survey report of relevant studies, note that a high degree of task dependency is evident in the patterns of performance on work sample ('strong') tests. They conclude that this dependency severely limits the degree of generalisability both across tasks and from the overall results of such tests to other supposedly similar tests. There is thus a tension between domain coverage and task-based assessment.

The issue of task dependency is obviously crucial to work sample type tests such as the PTEEP tests, where the aim of the testing is to tap the kinds of higher-order cognitive skills that are believed necessary for successful performance in Higher Education. Task-based testing, which sets up simulations of the 'ill-structured' problems (Strohm-Kitchener 1983) typical of the context, is arguably an appropriate approach to eliciting these higher-order cognitive skills. That is to say, higher-order skills, as has been argued earlier in this study, are most likely to be revealed through active engagement in problem-solving activities that require expertise and real deployment of skills. Tests that provide opportunities for such engagement, for example by including simulations of the kinds of extended problems encountered in Higher Education studies, are therefore most appropriate. In order to deal with the problem of generalisability, and to lessen the impact of task dependence, several test developers (e.g. Weir 1990, Hughes 1989, Bachman 1990, Bachman & Palmer 1996) have opted to specify micro-, or enabling, skills within these tasks, and then to investigate performance at both the task and the micro-skill level.

An attempt to investigate, albeit indirectly, the task dependency versus domain coverage issue can be seen in the development of the Tertiary Education Linkages Project (TELP) test in academic literacy, currently being undertaken by the author. This test, designed for use as a diagnostic tool

in the seventeen HDIs in South Africa, is based on a similar construct to that of the PTEEP. Because of particularly severe resource constraints, the inclusion of constructed response items, with their consequential requirements in respect of marking, could only be justified if it could be demonstrated that they elicit significantly different responses from candidates than do multiple-choice item formats.

For the testing of language, however, total reliance on multiple-choice formats tends to be strongly resisted by language educators, who challenge its legitimacy. Constructed response items require candidates to 'produce' language: that is, to write as compared to select from given alternatives as is the case with multiple-choice formats. Judging a student's ability to produce language (to write essays, laboratory reports, describe and construct graphs, etc.) from her/his ability to select the correct response from a range of pre-determined alternatives is clearly very difficult to justify without empirical evidence demonstrating a clear and strong link between the two (multiple-choice items and constructed response formats)[64].

A correlational analysis was therefore conducted of student performance in order to determine whether or not a student's performance on the multiple-choice items in the test relates strongly enough to her/his performance on the total of the constructed response items to suggest that the multiple-choice question (MCQ) total could be used alone, or as a proxy for constructed response performance. The results suggest that the MCQ total can indeed be used alone, as the correlation between the total of the MCQs and the total of the constructed response items was shown to be 0.81. This correlation is strong, and one can deduce from it that there are strong similarities between performance on the two format types. This deduction formed the basis for the decision to recommend a two-phase approach to the marking, where the multiple-choice and proof-reading items would be marked for all candidates, and the remaining constructed response items only for those candidates placed on remedial courses[65] (thereby providing important diagnostic

[64] The constructed response items in this test include creative writing, persuasive extended writing, correction of errors, and a proof-reading task. The multiple-choice items are set to cover the full-range of specifications – that is, to achieve construct representation.

[65] The proportion of candidates recommended for placement onto some kind of remedial programme varies according to discipline and institution. As a rule of thumb, only about a fifth to a quarter of all candidates would normally be expected to fall into this category.

information). This action would considerably cut costs, while preserving face validity and providing important diagnostic information where it is needed most (Yeld 2000).

In terms of the task dependency issue – that is, whether tasks constrain generalisability - it can be seen that in this test, candidate performance on the larger tasks correlates highly with their performance on items testing a range of skills and abilities[66]. It can therefore be concluded that in this context, task dependency is not a major issue.

When a similar analysis was conducted on the PTEEP tests (illustrated in Table 9.1 below), a relatively strong correlation was found between performance on the constructed response items as a whole, and the test total (0.7). The relationship between the constructed response items and the multiple-choice items, however, was only moderately strong (0.61). Combining the multiple-choice items and the proof-reading cloze item, which had proved a very useful combination in the TELP example above, yielded a correlation with constructed response of $r = 0.8$, suggesting that a similar range of abilities is being tapped in both types of items. This is an important finding for the PTEEP, as it was for the TELP project, as it provides support for continuation and possibly expansion of multiple-choice and cloze format use – a highly attractive option in a context of scarce financial resources.

	Constructed Response /120	Proof-Reading /20	MCQ + Proofreading /39	Total /159
MCQ (/19)	0.61	0.43	0.77	0.7
Constructed Response		0.73	0.8	0.99
Total	0.99	0.72	0.84	

Table 9.1: Correlations between Item Types in the 1998 PTEEP: (n=914)($p < .05$)

In terms of the tension between the extensive domain coverage made possible by a multiplicity of multiple-choice items, and the less flexible task-based tests, it seems from both the TELP and

[66] This is echoed in research conducted recently on medical students' reasoning on multiple-choice versus equivalent constructed-response items, which suggests that "... when items are well constructed and call for specific declarative knowledge that the students possess, it seems to make little difference which format is used ..."(Skakun and Maguire 2000: 14). When students do not possess the specific declarative knowledge, however, the authors suggest that different thought processes are used, with the MC distractors providing "major focal points for memory searches". Interestingly, the authors conclude that the presence of possible alternatives serves an appropriate and useful function by aiding the progress of the students from novice to expert. This function has been harnessed to some extent in the development of the PTEEP tests, where MC formats are used to provide quick and guided opportunities for candidates to reflect on the information provided in the texts, and to prepare for major tasks such as essay writing.

systematic approach. This preference is based on the argument that an instrument removes the tendency for a group to arrive at consensus on construct-irrelevant grounds: that is, on the basis of the dynamic the group has built up, where the opinions of certain members might carry more weight than others by force, for example, of seniority or personality, rather than by force of principle. A common sense approach to this problem might be to use both methods in combination, with the panel of experts using the outcomes of the collection instrument as a basis for discussion.

In the development of the PTEEP tests, the following procedure has been adopted. First, the principal test developer (the author of this study) writes a comprehensive report on the development of the previous test. In addition, a theme is selected from a number of topics identified as potentially interesting during the previous test development cycle, and a variety of possible texts on the topic is gathered. Copies of all of these are made and assembled into workshop packs for the test development group.

The second stage in the PTEEP development cycle takes the form of a test development workshop, usually lasting two full days. The first session in each of these workshops is set aside for discussion and analysis of:

- the test specifications. The specifications are analysed at this stage without reference to past or future tests, and the discussion highlights gaps in the specifications, new emphases[67], new understandings, and suchlike. It serves to assist in the induction of new members, and to further develop the common understandings of the group.
- the test development report of the previous test (including item analysis statistics). The discussion here focuses on analysis of the items, and is an effective way of drawing attention to problematic and promising item types, and to the possible causes for poorly functioning items: that is, items that do not discriminate effectively between strong and weak candidates.

[67] For example, over the years 1997 to 1999 in particular, 'understanding basic numerical concepts and/or information used in text ..." gained more prominence than it had previously been accorded.

PTEEP examples that it is possible to attain a degree of agreement between the two format types, in terms of what is being tested. However, it is important that this type of agreement should be monitored, not only statistically as shown above, but also through the use of expert judgements, as discussed below in section 9.2.

In the main, there are two categories of primary concern in relation to content validity: (i) construct representation, and (ii) construct relevance (Messick 1994).

9.2 Content Validity in Terms of Construct Representation

Claims for content validity on the grounds of construct representivity are based on the extent to which the content of a test constitutes a representative sample of the construct requirements. The focus in this regard is on coverage of the specifications, and Table 9.2 below illustrates one way of checking for this coverage.

The investigation of content validity frequently takes the form of a panel of experts comparing the items and tasks within a test with the domain specification of the test, and, where available, with performance statistics. One of the challenges in this process is that, for a test of academic literacy, the panel of experts should ideally be drawn from a wide range of disciplines. The panel could, for example, be comprised of curriculum development specialists, cognitive psychologists, social scientists, language development practitioners, science educationists, and assessment specialists. It is evident that a considerable amount of training is required in order for such a group to undertake the task within a common framework of reference. However, as McNamara (1996) cautions, such training could result in false conformity – the "risk of cloning", as Alderson et al (1995:175) put it.

A different approach to content validation involves the use of a data collection instrument, which could take the form of a rating scale of some kind. The scale could be based on degrees of complexity, numbers of occurrences, or some other feature thought relevant to the test construction. The use of a data collection instrument rather than a discussion conducted by domain experts is considered by some (e.g. Alderson et al 1995) to be a more objective and

During the remainder of the test development workshop, group members work in sub-groups, with language specialists paired with science educators and/or curriculum specialists. The focus is on generating items (multiple-choice and constructed response) on the texts which have by this stage been provisionally selected. At the end of the test development weekend, each sub-group presents its items to the whole group, where they undergo critique. It is at this stage, too, that the specifications re-surface, and gaps in coverage are identified and addressed.

The third stage of the development cycle is the responsibility of the principal test developer, who collates all the items and texts, and undertakes the laborious process of redevelopment of the items and where necessary, the modification of the texts so that they form a coherent whole in terms of the theme, and provide comprehensive coverage of the specifications.

Once the test has achieved first draft status, a small-scale pilot of the test is conducted, using first-year student volunteers. As the sampling is very opportunistic, no reliance is placed on levels of performance, but the process does reveal gross misunderstandings and confusions with items, rubrics and texts, and is helpful also in revealing how much time is needed to write the test.

After the pilot, the test development group meets again and critiques the draft, paying particular attention to coverage of the specifications (construct representation) and to the range and variety of items used. The results of the small-scale piloting of the test are discussed, and the test is modified in light of this information. After this modification, it is once more the responsibility of the principal test developer, who takes the test to its final form.

Table 9.2 below illustrates the extent to which the construct, as embodied in the specifications in the left-hand column, can be considered to be represented by the items and texts (the content) in the PTEEP tests from 1997 to 1999.

SKILL AREA	Items Tapping Skill Area		
	1997 'Antarctic' PTEEP	1998 'Fire' PTEEP	1999 'Water' PTEEP
Detailed reading for meaning: •	1.2, 1.10, 1.11, 1.12 2.1, 2.2, 2.4i-ii, 2.5, 3.1, 3.2, 3.3, 3.6, 3.7, 5.1, 5.3, 5.4, 5.5	6.2, 6.1, 1, 4.1, 7.1, 7.2, 9, 16, 18, 19, 20, 21.1, 21.2, 22.1, 22.2, 23, 33	1, 3, 4, 6, 9, 10.1, 11, 12, 13, 14.2.1, 14.2.2, 14.3, 15, 16, 17.1, 21, 22.1, 22.2, 25, 26, 28, 29, 30, 31.1, 33, 37
Skimming and scanning Using layout to obtain overview, surveying to obtain gist, scanning for specifics	1.1, 1.6, 1.10, 1.11, 1.12, 2.1, 2.2, 2.3, 2.5, 3.1, 5.3, 5.4, 5.7, 6.1, 6.3	1, 4.1, 6.2, 8, 9, 13.1, 13.2, 14.1, 14.2, 22.1, 22.2, 26, 28, 29	1, 2, 3, 4, 6, 10.1, 11, 15, 18, 22.1, 22.2, 24, 5, 28, 30, 31.1, 32, 33, 38
Vocabulary:	1.1, 1.2, 1.3, 2.4, 5.2	1, 6.1, 6.3, 17, 20, 28	9, 12, 13.2, 26.1, 26.2
Metaphorical expression	1.1, 2.4i, 3.5	6.2	17.1, 24.1, 24.2, 24.3
Extrapolation and application: (drawing conclusions, or applying insights, derived from texts, seeing trends)	1.7, 1.10, 2.2, 2.5, 3.1, 4.2, 6.3,	1, 2.3, 3.1, 3.2, 3.3, 4.1, 4.2, 4.3.2, 9, 10.1, 10.2, 11, 14.1, 14.2, 26, 28	5.5.1, 5.5.2, 8, 14.2.2, 22.1, 27, 33.2, 35, 37, 40
Inferencing: (understanding ideas/information. in a text, implied but not explicitly stated)	1.10,3.8,5.2	2.4, 4.3, 10.1, 10.2, 13.1,15, 21.1, 21.2, 24	18, 24.1, 24.2, 24.3, 8,25, 26.1, 26.2, 27, 30
Relations between parts of text through devices of cohesion (e.g. pronoun reference, particularly demonstratives, referring to statements/propositions or 'entities')	2.5, 3.3, 5.5, 5.6, 6.2, 5.5	7.1, 7.2, 12, 16, 18, 33, 35	14.1, 14.2.1, 14.3.1, 14.3.2, 14.3.3, 17.1, 21
Rhetorical devices such as those used in definition, exemplification, exhortation, argument/persuasion	2.2, 2.3, 5.1	12, 17, 35	15, 33.2, 35
Relations between parts of text and/or text extracts by recognising indicators in discourse, especially for: introducing, developing, transition and conclusion of ideas; signalling relations between phenomena	2.2, 3.1, 5.6, 6.3, 2.3, 2.5, 1.7	4.1, 8, 14.1, 14.2, 18, 22.1, 25, 26, 33, 35	6, 14.2.2, 15, 21, 22.1, 22.2, 32
The grammatical/syntactical basis of the English language	2.2, 2.3, 3.1, 3.2, 5.7, 6.3	4.1, 9, 12, 18, 27.3, 33, 35	11, 15, 21, 35, 40
Text genre (including audience, purpose etc.)	5.1, 6.3	12, 15, 17, 20.1, 20.2, 21, 30, 35	18, 31.1, 31.2, 39, 40
Information presented visually (graphs, tables, diagrams, maps, flow-charts)	1.9, 4.1, 4.2, 4.3, 4.4, 4.5, 1.7, 1.8	2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 5.1, 5.2, 5.3, 14.1, 14.2, 25, 27, 31.1-.4, 32.1, 32.2, 34	5.1, 5.2, 20.1, 20.2, 20.3
Separating the essential from the non -essential: • main idea from supporting detail • statement from example • fact from opinion • proposition from its argument • classifying and categorizing	• 3.6, 2.7, 5.7, 6.2 • 1.11, 2.2, 2.4ii, 3.2, 3.6 • • 1.11, 2.3 • 3.7	• 3.1-.4, 8, 12, 19.1, 19.2, 22.1, 35 • 13.1, 13.2 • 19 • 2.3, 2.4, 31.1 • 2.4	• 10.1, 13.1, 15, 16, 22.1, 22.2, 32 • 8, 13.1, 15 • 10.2, 26.1, 26.2 • 10.1, 18, 35, 40 • 7, 19, 35, 40
Basic numerical concepts and/or information used in text, basic numerical manipulations: • estimation • comparisons, greater than, smaller than • percentages, basic fractions (half of etc.) • basic chronological references, sequencing • basic computations	• 1.4, 1.9, 4.1, 4.3 • 1.6, 1.8, 1.9, 4.1 • 1.4, 4.2 • 3.5 • 1.5, 4.5	• 2.1, 2.2, 15, 34 • 5.3 • 2.2, 5.2, 11 • 15, 31.1, 26, 31.2, 34 • 5.1, 5.2,11, 27, 31.1, 31.4	• 5.3 • 5.3,20.3 • 36 • 29 • 2, 5.2, 5.4, 5.5.1-.5, 20.1-.3, 29, 3.1, 33.3, 34, 36
The importance of 'own voice' (including 'ownership' of ideas) and/or creativity of thought and expression	6.3	5.1, 27, 35	35
Conventions in text (e.g. referencing and references, footnotes,	3.4, 5.1, 6.2	17, 21	3, 39

Table 9.2: Coverage of PTEEP Language Knowledge Specifications 1997 - 1999

It can be seen in the table that extensive coverage of the specifications was achieved in each of the three tests. This is not to say that the coverage was identical in each test, but that a reasonable coverage was achieved in each case. Given this comprehensive coverage, it can be argued that the PTEEP tests are valid insofar as construct representation is concerned.

9.3 Content Validity in Terms of Construct Relevance

Investigations of content validity on the grounds of construct relevance set out to assess the extent to which items introduce variables unrelated to the test construct. One example of this intrusion can be seen in an English Second Language test developed in the 1980s by the Human Science Research Council in South Africa. The test, advertised as suitable for selection and admissions purposes for all South African ESL students, included a large number of questions on the meaning of such phrases as: "And you lot want to string the fives on Tojo just because he's got a Blighty fag case ", "Show him a rice pudding and he gets the screaming ad-dabs", or "... a sort of military Al Capone ...". Students who do not understand the meaning of such phrases would clearly perform poorly on the test – but it is difficult to make connections between this and their future performance at an English-medium South African university. The kinds of knowledge tapped by these items would be considered to represent construct irrelevance. Another, less extreme, example relates to the impact of language in a Science test. Some test developers see this impact as a source of construct irrelevance, whereas others argue that in the criterion situation, learners will encounter and have to deal with science concepts through the medium of language, and so it is a relevant aspect of the construct (e.g. Haeck et al 1997, Zaaiman 1998).

In order to guard against this kind of problem, it is essential to undertake what have become known as 'fairness reviews' (Norton Peirce 1992, ETS 2000). These can involve both detailed scrutiny of items by panels chosen on the grounds of diversity (e.g. relating to such grounds as sex, ethnicity, and/or religion) and statistically based investigations of how designated groups fare on items. In the PTEEP development process, the former can be said to be undertaken to some extent, as the development teams are constituted at least partly on the grounds of diversity. Because of the method employed in reporting the results and ranking the students, no systematic attention has thus far been paid to investigating bias. Current analyses reveal wide differences between performance of different groups (see, for example, Table 2.2, Chapter 2). However, these performance differences are ascribed to educational background, and are seen as an attribution of

cause[68] issue, rather than as one of test bias. Since groups are treated differently on the grounds of educational background (they are ranked separately on this basis), test bias has not been fully investigated. Future development cycles should, however, undertake appropriate studies in this regard. For example, use could be made of Differential Item Functioning (DIF) analysis techniques, which assess whether individual test items or tasks favour one group over another, where the ability of the groups is the same.

In terms of content validity, the requirement of construct relevance clearly means that the content of a test must be defensible in terms of its underlying construct. This, however, requires decisions to be made about how to operationalise the construct. According to Bachman and Palmer (1996:95), "[L]anguage use can be viewed as the performance of a set of interrelated language use tasks". They coin the term 'ability-task' to reflect their view of the inseparability of the two, and to highlight the role that language use tasks play in the mobilisation of language ability. It is to this role, above all else, that content validity must address itself. Provision of appropriate texts, for example, will be of little use if the tasks within the tests do not serve to provide appropriate opportunities for the candidates to demonstrate their ability in meaningful ways.

Clearly, the language used in the test tasks, including the texts in the test, and in the responses of the test takers to the tasks, must resemble the language used in the criterion situation. For example, the inclusion in a test of highly creative and imaginative instances of language use, or of tasks such as poetry writing or caption writing, would be inappropriate for a test designed to predict future performance in first year university courses - unless, of course, the candidates were intending to pursue courses in fields that placed a premium on such uses of language. Essentially, the challenge is to provide a range of academically situated texts and tasks that allow candidates to demonstrate their competencies and preparedness for academic study.

Ideally, these texts and tasks should be empirically derived and validated for specific contexts. An example of how this could be achieved can be seen in Weir's three-phase investigation into the

[68] Attribution of cause is discussed in Chapter Three, section 3.4.2.

language-related needs of overseas students studying in Britain (Weir 1983)[69]. The first phase of his investigation involved an extensive observation of Science, Engineering and Social Science courses at three British universities and three colleges. The findings from these observations were used to guide the development of a series of questionnaires, which formed the second phase of his investigation. The respondents - academics, British students and overseas students - were asked to estimate both the frequency and the level of difficulty of certain tasks that the observation phase had indicated were significant. The results of the questionnaire phase were used to identify, for the main skill areas (reading, writing, listening, and speaking), those key activities that were likely to pose problems for overseas students when studying at British universities. The third phase of Weir's research entailed the development and trialling of various test formats to assess which were best suited to eliciting the kinds of tasks and difficulties identified in the second phase.

Despite this exhaustive empirically based needs analysis, however, the "... end product was relatively unmanageable for purposes of test construction" (Alderson 1988:223). As Weir himself acknowledges, the type of needs analysis used in his investigation (based on Munby's (1978) Communicative Needs Processor) is "... unable to specify the relative importance of the variables" (Weir 1990:16). In addition, it is not helpful in specifying the degree of complexity involved in various instantiations of tasks and activities.

Interestingly, Weir's research findings did not support the development of subject specific versions of the Test in English for Educational Purposes. This finding was confirmed in the validation study of another large-scale British test, the English Language Testing Service (ELTS) test. Despite the lack of research evidence to support the validity of subject-specific components (i.e. showing that students are disadvantaged by taking a test not in their subject area), market-related factors indicated the wisdom of retaining them (Alderson & Clapham 1992). Currently, however, the IELTS tests do not contain subject-specific components. In the development of the PTEEP tests, it was decided, in light of severe resource constraints as well as the research findings noted above, not to develop subject-specific modules and/or versions. Nevertheless, in an attempt to ensure

[69] Further examples can be found in the work of Hale, Taylor, Bridgeman, Carson, Kroll and Kantor (1996), Cumming, Kantor, Powers, Santos and Taylor (2000), Ginther and Grant (1996), and Horowitz (1986).

even and fair coverage of topics and tasks, care was taken in composing the test development teams over the years so that a range of disciplinary perspectives was included at the design stage. Thus, the test development team included specialists from the following areas of expertise: curriculum development, science education, language development, assessment, social science, language arts, and admissions.

Given the resources allocated in the initial stages of the AARP initiative (see Chapter Seven), it was not possible to embark on an empirically based needs analysis exercise. Instead, the findings of Weir's comprehensive undertaking, in particular, were used as a starting point for test development. As the test was designed to be capable of being written 'under a tree', no reliance was made on technologies such as tape recorders, computers, or on test formats that required a high degree of contextual replicability (e.g. interviews, listening or speaking tests). In consequence, no direct attempt was made to test listening or speaking, or to test the following purposes (Halliday 1973) for which language is used:

- instrumental (using language to organise one's life - e.g. keeping a diary accurately, noting call numbers for library books or web page addresses for other sources of information);
- personal (expressing emotions and personal responses);
- interactional (phatic and social uses of language); and
- imaginative (the use of language to create new and novel meanings e.g. poetic).

The constructs of reading and writing were discussed in Chapter Six. The main purposes for reading in academic contexts were identified as follows (Enright et al, 2000): to find information; for basic comprehension; to learn; and to integrate information across multiple texts. These purposes were further broken down into what Weir (1983, 1990) and others have called 'enabling skills' – that is, the kinds of skills and abilities needed in order to successfully read for the purposes listed above. The texts within the PTEEP were selected with these purposes in mind, as is evident in section 9.3.1.2 below.

Cumming, Kantor, Powers, Santos and Taylor (2000), in their analysis of academic writing issues, identify core rhetorical functions associated with the writing of expository text, and suggest that effective writing in academic contexts depends on mastery of the following rhetorical functions: categorisation of key features of text, and their analysis in some way (e.g. by describing relations between them); problem identification, analysis and suggested response (solution); and stating of a position, elaborating it, and/or justifying it. These functions, of course, include activities such as describing, defining, enumerating, and comparing/contrasting, but these are normally used in the completion of broader tasks. These broad categories are elaborated somewhat by Horowitz (1986), who suggests that the writing demands most commonly encountered by students in higher education include synthesis of various sources; connection of theory and data; summary of/reaction to a reading; report on a specific participatory experience (e.g. experiment, field trip); research project; and annotated bibliography.

Taken together, these sets of purposes for writing fall into what Halliday (1973) has called heuristic (using language to learn) and representational (using language to convey facts and knowledge) purposes for language use. These two purposes correspond to the Bachman and Palmer (1996) categories of heuristic and ideational language use, which define the main purposes for which language is used in academic contexts.

Hale et al (1996) developed the following classification scheme after examining assigned writing tasks collected from eight universities in the United States and Canada. Six disciplines were represented at the undergraduate level, viz. Business, Chemistry, Civil Engineering, Computer Science, English and Psychology. They concluded that the kinds of writing tasks (genres) usually encountered in higher education are as follows:

Genre	Rhetorical Task	Pattern of Exposition
Essay Library research paper Report of experiment Observation with interpretation Summary (including annotated bibliography without comment) Case study Plan/proposal Documented computer programme Book/article review Unstructured writing (free writing, journal entries, notes)	Narration Description Exposition Argument	Process Classification/Enumeration Exemplification/Illustration Comparison/Contrast Cause-effect/Problem solution Definition Analysis

Table 9.3: Task Classification Scheme (Hale et al 1996)

Other features of this classification scheme include an attempt to define the "... level of thinking skills or intellectual functioning required to accomplish certain tasks" (Hale et al 1996: 12). This attempt was based on Bloom et al's (1956) taxonomy of educational objectives, and resulted in a two-part distinction: tasks that require candidates to 'retrieve' and 'organise'; and those that require candidates to 'apply/analyse/synthesise/evaluate'. These tasks were, respectively, categorised as representing 'lower-level' and "higher-level' cognitive activities. However, attempts to classify levels of cognitive complexity and demand are extremely difficult (e.g. Frederiksen 1994). For example, 'retrieving' information from a complex text can be a far more difficult undertaking than 'applying' information gained from a straightforward, simple text, to a similar situation. Nevertheless, the distinction between lower- and higher-level cognitive activities is useful even if only at the level of commonsense, and as a reminder, in the test development process, to consider cognitive task demands. This has not been done in any systematic way in the PTEEP development process to date, and it is recommended that future PTEEP test development procedures incorporate attempts to understand and define this important aspect of task demand.

Various kinds of extended writing are required in the PTEEP tests. They are shown in Table 9.4 below in relation to relevant sections of the Hale et al (1996) classification scheme.

Year and Theme of Test	Writing Task (title)	Genre	Rhetorical Task	Pattern of Exposition
1996 Education	The authors of the two articles have very different views about education. In a brief essay of not more than one page, compare and contrast the arguments of these authors about the value of competition in education.	Essay	Exposition	Classification/ Enumeration Comparison/ Contrast
	Drawing on the article as a whole, describe in not more than 10 lines what the author calls the 'phenomenon of qualification-getting'.	Summary	Esposition/ Description	Classification Cause-effect Analysis Definition
1997 The Antarctic	In a brief essay of not more than one page, outline the main arguments used in the texts to persuade readers about the need for a conservation strategy for the Antarctic. Your essay should make it clear to the reader why the Antarctic environment is in some danger, and why it should be protected – from the point of view of the world in general, as well as South Africa in particular.	Essay	Exposition/ Argument	Classification Exemplification Cause-effect Analysis Definition
	In not more than 100 words, write a summary of this passage (Explore Antarctica).	Summary	Exposition	Classification Analysis Cause-Effect
1998 Fire	Fire is often believed to be our enemy. Write an essay of not more than one page, in which you discuss the many ways in which fire has been (and continues to be) of benefit to the world. Much of the information you need is contained in the articles in this test booklet.	Essay	Exposition	Classification/ Enumeration Exemplification Definition Analysis
	Complete the following letter in the space provided, basing the content on the information in the article[70].	Summary	Exposition	Classification Exemplification Cause-effect Analysis
1999 Water	In approximately one page, outline the main arguments (put forward in the texts) against the commencement of Phase 1B [of the Lesotho Highlands Water Project]. The arguments should cover the disadvantages of the scheme for both Lesotho and South Africa.	Essay	Exposition/ Argument	Classification Exemplification Cause-effect Analysis Definition
	Write a letter to the Lesotho Daily News in which you point out the economic advantages for your community if Phase 1B does not go ahead.	Summary	Exposition/ Argument	Classification Cause-effect Analysis

Table 9.4: Extended Writing Tasks in PTEEP Tests, 1996 – 1999

The table reveals a narrow range of genres used in the tests. The reason for this is that candidates will have had very little exposure to extended writing tasks at school (Kapp 2000a&b, Department of Education 1998), and the genres included are the ones most likely to have been encountered. Similarly, relatively few expository patterns are required: however, the extensive and comprehensive coverage of underlying skills in these areas that is achieved in the multiple-choice and short-answer constructed response items can be assumed to provide coverage of these patterns.

[70] The article is a report and the letter is a summary of this. The format of the letter, and the opening sentence, are provided as part of the rubric.

In terms of construct relevance, the above analysis suggests that the test content is valid, in that the kinds of language and the choices of language task are clearly guided by and based on the test specifications and blueprint.

Deciding on the kind of language use task, however, is not sufficient. In order to help ensure construct relevance and representation, a framework is needed that will help to put flesh on the bones: how the construct should be operationalised needs to take into account the impact of the tasks themselves. In moving from specifications (which are properly thought of as part of the test construct) to test content – i.e. to operationalising the construct - many decisions need to be made. Jamieson et al (2000) developed a useful model as a framework to guide the development of the TOEFL 2000 project. The framework takes, as a starting point for the analysis of test content, the notion of 'task characteristics'. The rationale for this notion is that, irrespective of what the task is (i.e. whether it is a summary cloze, an essay, or a multiple-choice comprehension item), the ways in which it is and can be undertaken are influenced profoundly by its characteristics. Any examination of content validity must therefore of necessity involve an examination of task characteristics.

9.3.1 Task characteristics

The framework below represents a combination of those developed by Jamieson et al (2000), Bachman and Palmer (1996), Yeld and Haeck (1997) and Yeld et al (2000, 1999, 1998).

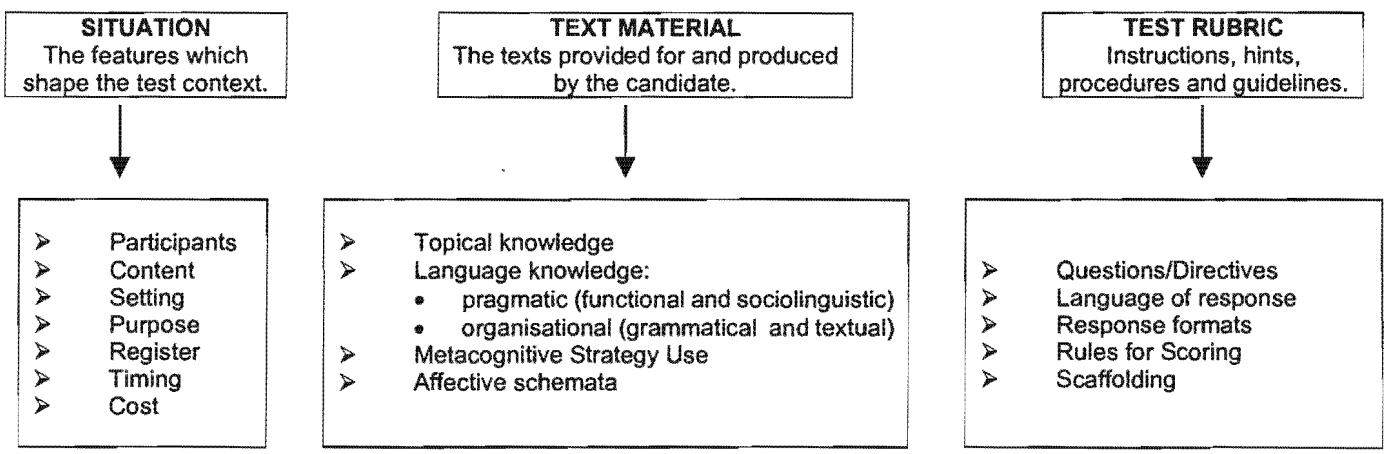

Figure 9.1: A Model of Task Characteristics

In sections 9.3.1.1 – 9.3.1.3 below, the three categories of task characteristic (respectively, situation, text material, and test rubric) are discussed in relation to the development of the PTEEP tests. The purpose of the discussion is to assess to what extent the PTEEP tests can be considered to be valid in relation to these characteristics.

9.3.1.1 Situation

'Situation' refers to extralinguistic features of the testing environment that impact on test performance. Assigning importance to this set of task characteristics is in accord with views of communicative language ability that highlight the importance of context for language use, as well as of situative and cognitive notions of knowing and learning which emphasise the roles of factors beyond the learning task, such as the connections between learning tasks, and learners, in learning situations. These views and notions are discussed in Chapters Five and Six in particular, and are not repeated here.

As can be seen in Figure 9.1 above, the task characteristic of situation comprises a number of elements, namely participants, content, setting, purpose, register, timing and cost. The discussion below focuses on whether the PTEEP tests can be said to be valid in terms of these elements. Clearly, validity in these terms is not possible to assess statistically: for this reason, the discussion below aims to make clear the extent to which the PTEEP tests have taken seriously the impact of situational elements on test performance, rather than to be able to state categorically to what extent the tests are valid in respect of each of these elements.

For a large-scale pencil-and-paper test such as the PTEEP, which targets reading and writing, 'participants' are not as crucial as they would be, for example, in a test of speaking based on an interview. The decontexualisation of academic language use described by Cummins (200,1984), Snow et al (1991), Tannen (1985) and others is relevant in this context. Nevertheless, it can be argued that, in reading, the reader, the authors of the texts, and in some sense the arguments or events in the texts, can be regarded as participants in the test-taking situation. In the development of the PTEEP, care was taken with such aspects as the choice of texts and themes, to ensure that particular groups are not excluded. In addition, invigilators (often unacknowledged participants in

the testing context) are briefed to be helpful and encouraging, as they are influential in setting the 'tone' in the examination setting.

'Content' is another situation aspect of considerable significance. As was discussed above in relation to Weir's needs analysis research, the jury is still out on the relative merits of discipline-specific versus non discipline-specific topics on test performance (Alderson et al 1995). That is to say, it is not yet been shown that candidates are disadvantaged by taking a test not in their broad subject area. Until this issue has been clarified, it makes sense in a context of limited resources to select subject matter that is as general as possible in order to minimise the number of test versions needed. In the development of the PTEEP, the construct definition includes topical knowledge in the sense that topical information of increasing complexity is provided in order to provide a basis for the demonstration of ability. This finds form in the PTEEP tests in the provision of a theme which is intentionally cross-disciplinary. For example, the Antarctic theme of the 1997 PTEEP contains texts that cover topics such as conservation, tourism, and various social issues relating to the use of scarce resources.

According to Jamieson et al (2000), setting, the third aspect of situation, is "... the place where the language act occurs". For testing, the implication is that the test setting must include a range of settings, which represent the places where language acts might occur in the criterion situation. In the TOEFL 2000 development project (the context of the Jamieson research), the settings proposed for the language tasks include three broad types; instructional milieu (settings for formal instruction such as lecture rooms, or laboratories); academic milieu (non-formal academic spaces such as libraries, writing and computer centres); and non-academic milieu such as the financial aid office, the student health centre, or the residences. The TOEFL test, however, includes spoken and listening components, and thus can more authentically incorporate a wide range of settings than can a pencil-and-paper test like the PTEEP. In the development of the PTEEP tests, the language use domain is predominantly that which would normally take place in formal instructional settings. In general, formal academic texts and other resource materials that are expository in nature are used.

Purpose, as an element of situation, refers not to the purpose of the test (which might be to select or place students, for example) but to the purposes for which students in the criterion situation need to use language. Many attempts at creating taxonomies of reasons for language use exist (e.g. Bachman 1990, Heath 1980): it is generally agreed, however, that the predominant purposes for which students use language in a tertiary environment are to convey and to create knowledge. In the development of the PTEEP, the purposes pinpointed for inclusion in tests are, in the main, heuristic and ideational. That is, candidates are expected to use the language in the tests to learn about the theme and to plan and execute their responses, and to convey their understandings in a variety of ways. These purposes are reflected in the PTEEP tests, as the examples given in this chapter in particular illustrate.

In terms of register, the PTEEP restricts itself to formal language use, as this is the most common form of written language students will encounter, and be required to produce, in the criterion situation.

Finally, the timing of the test is an important extra-linguistic variable affecting task performance. If the test is written at the end of a day, or after several other assessment measures, the candidates are likely to be fatigued and/or demoralised, for example. It is not always possible to control this situational aspect. The long distances travelled by many candidates to get to the PTEEP testing centres results in a certain amount of fatigue, but considerations of cost and security make it impossible to increase the numbers of testing centres. In terms of how long candidates have to write the test, the test is not 'timed', in that (on the basis of information gained during the trialling phase), adequate time is provided for the great majority of candidates to complete the test. However, there is a cut-off time, as experience has shown that without this, some candidates will write for hours to very little effect.

On the basis of the discussion and analysis above, the PTEEP tests can be argued to be valid in terms of, or at least to have made serious attempts to address the requirements of, the seven aspects of situation (the first component of the task characteristics model illustrated in Figure 9.1).

9.3.1.2 Text material

The term 'text material' here includes all the texts that form part of the test - that is, the articles, illustrations, graphs, and so forth. As used here, it does not include test rubric elements such as the items within the test. As the criterion situation of the PTEEP tests is that of university study, it is desirable that the text material in the tests correspond as closely as possible in terms of these aspects to those used in university courses. Identifying the aspects makes this correspondence easier to achieve, and makes discordances more easily visible. Both of these are important considerations in test design.

The four components which, in terms of the model in Figure 9.1 above, need to be addressed with respect to text material are topical knowledge, language knowledge, metacognitive strategy use, and affective schemata.

In Chapter Six, it was argued that topical knowledge[71], while inescapably present in any situation, can either be downplayed or highlighted in language tests, depending on the purpose of the test. In the PTEEP test construct, the role of topical knowledge is understood as a crucially important component in the task performance of candidates. However, not all candidates will have the same kinds or extents of background knowledge – indeed, in a context of educational inequality such as exists in South Africa, these differences are acute. Therefore, in order to include topical knowledge as an important component of language-based performance, but also to downplay as far as possible the differences in educational opportunities prior to the test, the tests deliberately set out to create topical knowledge via the use of a complex and elaborated theme. It could be added, however, that the importance attached to topical knowledge in the construct of the PTEEP tests, coupled with the necessity of providing this knowledge as far as is feasible within the test, puts a complex spin on the choice of texts. Every text needs to be examined in terms of its 'yield' – that is, to what extent the text adds to the thematic complexity of the test as a whole, and to how efficiently it provides maximum yield in terms of item construction and construct representation.

[71] That is, the information base constructed by an individual, comprising such kinds of knowledge as cultural and content-related knowledge.

The range of topics and kinds of texts included in the PTEEP tests have been discussed from various perspectives in this and previous chapters in this study (for example, the topics and texts are discussed extensively in Chapter Six, section 6.4.2, and Chapter Nine, section 9.3, in relation to the role of content and below, in relation to language knowledge. On the basis of these discussions, it is argued that the PTEEP tests can be considered to be valid, in terms of the PTEEP construct, in respect of their treatment of topical knowledge.

Language knowledge is the second component of text material included in the model of task characteristics (Figure 9.1). In terms of grammatical features, attention needs to be paid to such matters as readability scores, sentence types, distribution of word classes and verb types, types of subordinate clauses, word frequency, semantic characteristics, and register features. For instance, a short text that uses a very high proportion of simple sentences with no embedded clauses and few abstract words, is unlikely to represent the same level of challenge for candidates when used as the basis for a task as would a text more representative of those used for university level study. However, the interaction of topical knowledge, task mediation and perceived complexity of text is complex, and each text thus needs to be assessed in context. Regarding pragmatic features of texts, as the main purposes for which students use language are heuristic and ideational, it follows that the texts within a test designed to tap these purposes, should lend themselves appropriately to this design. That is to say, the texts should, taken together, provide a range of types that correspond to those which are encountered by students in the criterion situation. Discourse features include rhetorical properties and text structure properties. In terms of the former kind, texts can be classified as (for example) descriptive, classificatory, contrastive/comparative, narrative, or argumentative/persuasive. In many undergraduate contexts, the "...emphasis of writing is often on telling people about the knowledge one has, or is acquiring, rather than on using writing to create unique or novel knowledge ..." (Cumming et al 2000:5, emphasis in original). The point is hereby made that tests should prioritise the inclusion of expository texts as the basis for tasks which will tend to involve the transmission and organisation of knowledge.

In the PTEEP tests, the texts are varied, although all the tests contain the types described below. The introductory text in each test is set out as a 'fact file' (Example A[72] below). While not a list, it contains simple sentences, is short, and uses straightforward description[73]. Additional texts in each test have the following characteristics: one is a medium length 'popular' piece of writing, usually adapted from a newspaper article, which is typically cause-effect in rhetorical type and, while using simple sentence and paragraph types, makes greater use of adjectives and persuasion (Example B below). A third kind of text included in each test is more complex, and is usually adapted from an academic source such as a journal (Example C below). In addition to these texts, all the tests contain a pie-graph, a line graph, a number line of some kind, and an assortment of tables.

Example A
SOME FACTS ABOUT FIRE
Fire is a simple chemical reaction involving oxygen from the air, materials which will burn, and a source of heat. Materials that are commonly burned for the heat they give (such as wood, coal or oil) are also known as fuels. The source of heat does not have to be an actual flame.

Example B
FIRE CHIEF FEARS INFERNO IN CITY
Hundreds of Cape Town's public buildings are potential fire hazards, as more than 80 percent of them do not comply with modern fire safety rules, the city's fire service has disclosed. But the department was unable to force property owners to safeguard their buildings, said Neville Evans, senior divisional fire prevention officer.

Example C
OLD FLAME
It is very difficult to prove that early humans deliberately used fire, particularly to show that they exercised some measure of control over it. Examples of control could include feeding a fire with fuel, attempting to spread or carry it, or confining and maintaining it in a hearth. It seems likely that early humans took advantage of fires that occurred naturally – for example, they probably ate the cooked flesh of animals trapped in grass fires. Traces of fire that are found where there are early human remains do not allow us to drew any automatic conclusions about the deliberate use of fire by early humans, however.

With respect to the text material component of language knowledge, the PTEEP tests, on the basis of the discussion above, can be argued to be valid, in that they incorporate an appropriate range of relevant text types, which incorporate appropriate grammatical, discourse, and structural features and elements.

Strategic competence, or metacognitive strategy use, is the third component of text material in the task characteristics model. The PTEEP construct (Chapter Eight, section 8.2.2) states that the

[72] The examples that follow are all taken from the 1996 PTEEP, which used the theme of 'fire'. In each case the extract is the first paragraph of the text, accompanied by the title of the text.
[73] The fact file texts are selected to yield reading ease scores which place them in grade levels equivalent to senior secondary schooling, following the Flesch Readability Formula (Flesch 1948).

tests will "not directly assess strategic competence, but will aim to ensure its inclusion in the design". In other words, the challenge for content validity is to ensure that the items and tasks in the tests create opportunities that will require the deployment of metacognitive strategy use. In the development of the PTEEP tests, it was decided to develop and include the kinds of tasks in whose completion the use of such strategies can be assumed, rather than to test them directly. That is, tasks would be developed which require students to plan and to monitor their performance. As is described below in Section 9.3.1.3, the rating scale for the essay task in the PTEEP tests includes a component for 'Organisation', which sets out criteria specifically for such phenomena as sequencing, logical and/or otherwise effective overall organisation. This rating scale is an example of what can arguably be regarded as an indirect measure of strategy use. Because the construct is only indirectly measured, however, no specific inferences can be made about candidates' use of these strategies, which it is believed will create conditions for metacognitive strategy use. A more direct probe into candidates' use of metacognitive strategies was outlined in Chapter Eight, Section 8.5.2.2, where an investigation into the extent to which candidates made use of a note-making task to prepare for an extended writing task was reported and analysed. The study concluded that stronger students appeared to be able to mobilise their metacognitive strategies while weaker students could not.

However, to date the test development process has not undertaken the kinds of research (such as think-aloud protocols) that would begin to yield systematic information on how effectively the tests do in fact call upon such competencies. It is recommended that future development cycles undertake such research. At this stage, however, it is not possible to assess, directly, the extent to which the PTEEP is valid in relation to metacognitive strategy use.

In connection with the impact of affective schemata, the PTEEP construct asserts that the tests will acknowledge the impact of affective schemata on performance, and will make efforts in the design and layout of the test, as well as in its administration, to minimise negative effects as well as to promote and harness positive effects" (Chapter Eight, section 8.2.2). In many ways, the scaffolding approach adopted in the PTEEP design addresses this issue, as the aim of the

scaffolding is to assist candidates to know what is expected of them, and to provide some guidance in the completion of the tasks. In addition to its other important functions, this is intended to reduce test anxiety. It is difficult to analyse with precision the extent to which the PTEEP tests have met the requirements of the test construct in this regard: that is, the extent to which the tests succeed in reducing, in whatever ways are feasible[74], the negative effects of a phenomenon such as test anxiety, and in harnessing positive aspects of affective schemata, such as motivation, perseverance and involvement. However, the evidence presented in Chapter Eight on the impact of scaffolding, as well as the design features described here, suggests that the tests can be argued to have taken seriously the issue of the role of affective features in test performance.

9.3.1.3 Test Rubric

Test rubric refers to those parts of the test that are not part of the thematic content. It includes the instructions given to candidates, the hints and procedures suggested in the test, and any guidelines that might be offered to assist the candidates in accomplishing the tasks. These elements play crucial roles, in different ways, in helping to determine the level of difficulty of tasks based on the texts. The example below is of a series of questions placed immediately after an essay prompt.

REMINDER

- **Have you used information from the readings? In particular, you should re-read:**
  - *** The article on page 2**
  - *** The summary on page 3**
  - *** The last article (on page 11)**
  - *** Your summary (see question 4 on page 13)**
  - *** You might also find your answer to question 11(ii) on page 10 useful.**
- **Have you used your own experience? Remember, only do this if you're sure it is relevant, and will illustrate or support a point you are making.**
- **Have you checked your work?**

The four aspects of test rubric in Figure 9.1 are questions/directives, response formats, language of response, rules for scoring, and the inclusion of scaffolding devices.

[74] These include such features as the layout of the tests (e.g. the use of illustrations and the employment of user-friendly language in the instructions), and the choice of themes, and the inclusion of a meaningful link to South African concerns.

In terms of questions and/or directives, it is clear that a great range of variables operates in determining the level of difficulty of an item. A question requiring the candidate to identify a concrete entity in a text (e.g. a person, a country, an event) is, on the surface at least, an easier task than one that requires identification of causes, effects, or outcomes. However, this inference does not always hold: some texts are constructed in such a way that identifying an entity is more difficult than a cause and effect task on a more straightforward text. Nevertheless, some form of hierarchy of difficulty is a useful starting point to guide test development in terms of questions and directives. In the PTEEP development cycle, as can be seen from Table 9.2, more than one item is developed to tap any one particular skill, and detailed attention is given to ensuring that a range of difficulty is included.

Response formats – the second aspect of test rubric - are self-evidently an important element of task characteristics. Despite this importance, however, research on how different response formats actually affect candidate performance does not reveal a straightforward picture. Although it is often assumed that a test which relies entirely on multiple-choice does not provide a valid basis for making inferences about writing ability, this assumption is not necessarily valid, as the examples in Section 9.1 (of the TELP and PTEEP tests) demonstrate. Moreover, neither is it clear that multiple choice and constructed response formats produce very different results for reading comprehension (Traub 1993). At this stage it seems that test developers are best advised to use as broad a range of response formats as possible, and the PTEEP tests have adopted this approach.

The language that is expected in the response is determined by the test specifications, which are themselves based on the test construct. Similarly, the methods employed in the scoring of responses are determined by the test construct. In particular, the construct defines the areas of language ability that need to be scored; the type of score to be recorded (e.g. overall score, profiles of scores according to skill area such as reading or writing); and the type of response format.

In the PTEEP tests, the extended pieces of writing are scored using a modified form of the Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey ESL Composition Profile (1981). Prior to 1987, a holistic nine-band score was used, with each band being characterised by several descriptors. However, severe difficulties were experienced in training scorers to use these bands, as well as in dealing with the high levels of anxiety generated by such common difficulties as assigning a band score to candidates who write relatively fluently but display little understanding of the materials they have read, or to candidates who have severe problems in communicating clearly but nevertheless display advanced reasoning skills. It was therefore decided to move to a rating scale system that separated content from writing. It is important to note that, although the PTEEP construct definition explicitly excludes topical knowledge outside of that contained in the text material, effective expository writing must be written about something – therefore the role of content needs to be acknowledged and addressed.

The attribute scale of Jacobs et al (1981) grew out of the need for an efficient and reliable means of evaluating the large number of compositions written as part of the English Proficiency Test (Michigan Test Battery). It is based on a system of weighted components, mastery levels and criteria descriptors, devised to focus the attention of readers from diverse academic and experiential backgrounds on important aspects of the writing samples, and to define the standards for each level. Selection and weighting of the components was based on extensive research, and it was thoroughly investigated in terms of reliability and validity in the context in which it was used. The weightings and categories of the Jacobs et al scale are as follows: Content (30%), Organisation (20%), Vocabulary (20%), Language Use (25%) and Mechanics (5%).

In terms of the language knowledge component of the Bachman and Palmer model of communicative language ability put forward in Chapter Six, these categories correspond, somewhat approximately, as follows:

Bachman and Palmer (1986)	Jacobs et al (1981)	SCORE RANGE (PTEEP)	CRITERIA
Functional	CONTENT	13 – 17	EXCELLENT TO VERY GOOD: knowledgeable (where relevant to task), substantive, thorough development of thesis, relevant to assigned topic.
		8 – 12	GOOD TO AVERAGE: Some knowledge of subject; adequate range; limited development of thesis; mostly relevant to topic, but lacks detail.
		4 – 7	FAIR TO POOR:: Shows only very limited understanding of topic, little substance, mostly irrelevant to topic
		0 - 3	VERY POOR Does not show knowledge of topic; non-substantive or not pertinent; OR not enough to evaluate.
Textual	ORGANISATION	8 – 9	EXCELLENT TO VERY GOOD: Fluent expression; ideas clearly stated/supported; succinct; well-organised; logical sequencing; cohesive
		6 – 7	GOOD TO AVERAGE: Somewhat choppy; loosely organised but main ideas stand out; limited support; logical but incomplete sequencing.
		3 – 5	FAIR TO POOR: Very choppy; ideas confused or disconnected.
		0 - 2	VERY POOR: Non-fluent; does not communicate; OR not enough to evaluate.
Sociolinguistic and functional	VOCABULARY	8 – 9	EXCELLENT TO VERY GOOD: Sophisticated range; effective word/idiom choice and usage; appropriate register
		6 – 7	GOOD TO AVERAGE: Adequate range; occasional errors of word form/idiom form, use etc. but meaning not obscured.
		3 – 5	FAIR TO POOR: Limited range, frequent errors of word/idiom form or use; meaning unclear because of errors
		0 - 2	VERY POOR: Lifted straight from text; OR not enough to evaluate.
Grammatical	LANGUAGE USE	12 – 13	EXCELLENT TO VERY GOOD: Effective, complex constructions; few errors of agreement, tense, number, word order/function, articles, prepositions, pronouns. Demonstrates mastery of conventions of punctuation, paragraphing, etc.
		9 – 11	GOOD TO AVERAGE: Effective but simple constructions; minor problems with more complex constructions; several errors in syntax but meaning not obscured or not often.
		5 – 8	FAIR TO POOR: Major problems in simple/complex constructions; frequent errors of negation, agreement, tense, run-ons, etc., meaning confused because of these.
		0 - 4	VERY POOR: Virtually no mastery of sentence construction rules, dominated by errors, OR not enough to evaluate
Grammatical and textual	MECHANICS	0 - 2	From demonstrating mastery of conventions (punctuation etc.) to frequent errors.
		50	TOTAL

Table 9.5: Adaptation of the Jacobs et al ESL Composition Profile (1981)

The Jacobs et al scoring scale is an 'annotated holistic scoring rubric' (Nitko 2001:196). That is, it specifies categories (content, organisation, vocabulary, language use, and mechanics) for which separate marks will be allocated, but within each of these categories, quality levels are defined. It

was suggested above that one of the main reasons for choosing to use an analytic scale, based on categories, is that it greatly reduces marker anxiety and makes moderation more transparent- that is, it facilitates a sharper focus.

As the table below demonstrates, there is a high degree of marker agreement when this scale is used. The data are derived from the 10% of the scripts (randomly selected) that are routinely double-marked. The following limitations on the data should be noted. First, the analysis is based on only one PTEEP test, the 1999 'Water' PTEEP. This is because the other PTEEP test results were not entered at this level of detail, and so the data are not available. Second, the analysis is based on two sets of marks for each of the individuals in the sample. However, the effects of different markers (as individuals and as pairs) are not examined. The emphasis is on the relationship between the first mark awarded to a candidate (by any marker) and the second mark awarded (by any other marker). Finally, the small number of candidates is a consequence of the fact that only the scripts of students who registered, as opposed to simply applying and not registering, were entered in this manner. Future investigations should be based on all moderated scripts.

1996 'Water' PTEEP	Content	Organisation	Vocabulary	Language Use	Mechanics	Essay Total	PTEEP total
Pearson's r	.89	.9	.88	.86	.52	.92	.99

Table 9.6: Correlation between First and Second Marking (1999 'Water' PTEEP) (p = 0.00) (n=61)

The category that stands out is that of Mechanics, which reveals relatively weak correlations. This component, however, was out of only 2 marks, and in fact is only appropriate for very low levels of language proficiency. Future PTEEP tests have increased the mark allocation to 3, taking one mark from the Content category, and this has improved the reliability of the marking of the Mechanics category (the 2001 test reveals a correlation of .76).

In terms of rules for scoring, it is argued that the PTEEP tests can be considered to be a valid operationalisation of the PTEEP construct. The variety of item formats, with their range of scoring approaches, and in particular the scoring scale used for the extended writing items (see Table 9.6), ensures comprehensive coverage of the test specifications.

The role of scaffolding as part of the test rubric is complex. As Bachman and Palmer (1996:50) note, test rubric features "provide the structure for particular test tasks ... that indicate how test takers are to proceed in accomplishing the tasks". In this sense, the kind of scaffolding that is used to alert candidates about how their work will be assessed forms part of the test rubric. An example of this can be seen in the PTEEP tests, where students are informed of the categories (which are themselves glossed) of the rating scale which will be used to assess their essays. This information serves as a reminder that they must pay attention to content, organisation, vocabulary and idiom, as well as language use. They are also advised to make notes before writing the essay, and a clearly labelled space is provided for the note-making task.

A different kind of essay preparation is illustrated in the example below.

Read the following statements and identify, in each case, from which source (which article) in the test the information comes.

1. Military activity in the Antarctic would have serious consequences for South Africa.

Title of Source (article)	
Introducing Antarctica	
Assessing the Costs to South Africa	
The Need for a Conservation Plan	
Explore Antarctica	

The purpose of this item (the example above was the first one of four) was twofold: to underscore various core concepts that would be useful in the essay task immediately following it, and to alert students to the need to acknowledge their sources. Whether the item is part of the text material or the test rubric, however, becomes a moot point: it is adding to the information by restating it, and it is helping to define the task. In the end, it probably does not matter much how it is categorised, except perhaps as a warning that if the intent of scaffolding is to help shape a task, it might be counterproductive to use it to add thematic information as this might increase the information processing load rather than assist the candidate.

In terms of test rubric, as discussed above, the PTEEP tests can be considered to be valid. The theory of language and academic performance on which the tests are based takes the role of sch

matters as instructions and scoring procedures very seriously: it requires, for example, that test tasks be very clearly and carefully constructed, including the use of scaffolding, in order to provide fair opportunities for candidates from a wide range of educational backgrounds. This results, as can be seen in the discussion above and in previous chapters, in the construction of a large variety of response formats, textual materials, and item types, which are included to reveal as full a picture as possible of an individual's ability to perform in the criterion situation.

Sections 9.2 and 9.3 above investigated the extent to which the PTEEP tests can be considered to be valid in terms of content validity. The analysis led to the conclusion that the tests are valid in several important respects, but that future development cycles should be encouraged to develop and introduce more systematic approaches to investigating content validity. Such approaches would require more comprehensive data entry practices, as well as the undertaking of various qualititative approaches to establishing content validity, including think-aloud protocols. Nevertheless, available evidence points to a considerable degree of content validity having been achieved.

9.4 Face validity

Face validity, as its name implies, refers to what a test looks like, and what it appears to measure. While there is no statistical measure for establishing face validity, it nevertheless remains a powerful factor in test use, influencing both test performance and test user confidence. One of the problems associated with face validity is that if test developers and/or users are not confident about conducting the statistical tests that are necessary in relation to the more technical validities, or in interpreting the results of such tests, it is likely that an undue reliance will be placed on face validity. As Stevenson suggests, this lack of technical expertise is a major reason for the "seductive appeal" (1985:112) of face validity. Some of the misconceptions arising from a disproportionate reliance on face validity are as follows (Yeld 1987):

- To be valid, tests must appear to be valid. This requirement is somewhat similar to the popular belief that to be effective, a medicine must taste unpleasant, or be costly.

- Tests that mirror as closely as possible a criterion situation are that situation. The danger here is that scores will not be regarded as simply the basis for prospective inferences from a simulated, sample situation to the criterion situation, but might be interpreted as actual instances of that situation.
- Validation studies are transferable across tests, regardless of such matters as populations, conditions, and scoring procedures, simply because two tests appear to be testing the same thing.

Notwithstanding these difficulties, face validity is an essential aspect of overall test validity. Not only, as suggested above, does it impact on candidate performance; it has a profound influence on the kinds of teaching and learning that precede it. The impact of assessment in this regard was discussed in Chapter Three, where it was concluded that while it is difficult to claim that tests can of themselves promote good teaching and learning practices, it is clear that bad or inappropriate tests can seriously distort these. If the test at the end of a period of learning is not perceived as reflecting the teaching that has taken place, it is likely that the subsequent teaching will suffer. As Davies (1985:4) points out, "... what the student's gaze (and the public's) is fixed on is the test, no matter how unreconstructed that may be".

Although face validity is not statistically assessed, it is possible to gather relevant data by interviewing students after they have written, or administering a questionnaire of some kind. It the gathered data are sufficiently detailed, it is sometimes possible to relate information to test performance – for example, when a particular group of candidates reports finding certain items alienating, or novel, it is possible to establish though correlational analysis, whether these perceptions are related to performance on these items. This retrospective analysis is useful for test development, as it bears on whether such items should survive in future versions, or be replaced. It is also useful to gather face validity information from test users, and not only test takers, as their perceptions about the test greatly influence the value they will place on the test results.

The initiatives that have been attempted in this regard – such as frequent presentations at Faculty Boards, where examples of the tests are discussed and displayed, or having Deans and admissions officers actually write sections of the tests – appear to have been successful in that feedback suggests widespread acceptance of and respect for the tests. However, it cannot be claimed that the PTEEP development project has undertaken systematic studies of the kind outlined above, and it is recommended that future cycles should do so.

9.5 Response Validity

Response validity is closely related to face validity, although the focus is more on ascertaining the reasons why candidates respond as they do, than on their perceptions of the test. The processes of reasoning engaged in by candidates are the targets of response validity data gathering, rather than their actual responses, although these elements are of course essential. The information gathered in this approach to validity is at the heart of what recent cognitive and situative theories of learning and knowing have foregrounded as central challenges for testing. An example of this recent perspective can be found in the National Academy of Education's final evaluation report on the National Assessment of Educational Progress project in the United States, which argues for a reconceptualisation of the NAEP assessment domains to include "... such aspects of student cognition as problem representation, the use of strategies and self-regulatory skills, and the formulation of explanations and interpretations" (National Academy of Education 1997, cited in Pellegrino et al 1999:124). Investigating how students approach and solve problems is viewed as just as important as whether or not they actually solved the problems.

Gathering such data involves either retrospective or concurrent introspective data collection procedures. The former procedure usually involves interviewing candidates after they have written a test, on why they have produced certain responses. This is frequently conducted on the basis of their actual responses. Concurrent introspective approaches involve such methods as 'think-aloud-protocols', where candidates are asked to articulate their thoughts and actions while actually taking the test. Clearly, this approach is not possible in situations where test performance is important, as candidates are unlikely to agree to be interviewed in such a situation – nevertheless,

it can yield important information in the test development process about what tests are really testing, as well as on the sometimes unexpected impact of new item formats. It is recommended that future development cycles undertake think-aloud protocols during the trialling of tests, and that the feasibility of interviewing students after they have written the test be investigated. It is possible that that would be feasible for students who write the tests after admission, when the candidates are registered on campus. Despite the problems of truncated samples which will arise in this situation, important insights are likely to be gained.

No systematic approach has been made to establish response validity in respect of the PTEEP tests. However, attempts have been made to gather feedback in the following ways: asking invigilators to speak to candidates about their perceptions of the test as they leave the exam room; giving questionnaires after the pilot of certain tests; and conducting regular information sessions for officials such as Admissions Officers. Results of these attempts, however, have been unsystematic and opportunistic, and have not provided the data on which validity assertions could be based. It is recommended that these shortcomings be addressed in future cycles.

9.6 Conclusion

In this chapter, the content, face and response validity of the PTEEP tests has been subjected to examination. In broad terms, the tests can, in light of the evidence, be considered to be valid, particularly in respect of content validity. Several recommendations were made about how future development cycles could strengthen the investigation of the tests' claims to all three aspects of validity. These include developing rating scales to use with the test development teams, data collection instruments to check on various aspects of face and response validity, and think-aloud protocols to gain insight into the cognitive demands of various items and item types. In addition, recommendations were made about tightening up on and expanding future data entry procedures so that properly designed research into the PTEEP tests could take place.

Chapter Ten turns to the crucial area of predictive validity, and analyses the extent to which the PTEEP tests can be said to provide significant and useful information about students' future academic progress.

CHAPTER TEN

PREDICTIVE VALIDITY
AND
THE PLACEMENT TEST IN ENGLISH FOR EDUCATIONAL PURPOSES

10.1 Introduction

In Chapters Eight and Nine, internal aspects of validity in relation to the PTEEP were discussed. The aspects were those of construct, content, face and response validity. It was argued that while in many instances the data were not sufficient (or had not been sufficiently systematically organised at the data collection stage) for firm conclusions to be drawn, there were persuasive indications that the PTEEP instruments met the basic criteria for validity in these respects.

Chapter Ten considers the predictive validity of the PTEEP tests; one of what Alderson et al (1995) and others have called external aspects of validity, viz. predictive, concurrent and consequential validity. These aspects, as the name implies, refer to how the test relates to phenomena external to the test, for example to other tests, to future performances in the 'real' world, and to educational and other systems in civil society.

10.2 Predictive validity

Approaches to establishing predictive validity focus on "... selected relationships with measures that are critical for a particular applied purpose in a specific applied setting" (Messick 1989:17). Admissions tests, which have as their aim the prediction of future performance in specified domains, are naturally particularly concerned about predictive validity. The aim of predictive studies is to predict future behaviours by examining the relationships between specified future behaviours and performances on a number of predictor variables.

Techniques for assessing predictive validity fall into two main groups. The first of these focuses on comparing scores on a test with an appropriate measure of performance in the target situation, and the second on comparing scores on a test with ratings of ability assigned by appropriate third parties.

Techniques within the first group frequently take the form of correlational studies. Such studies, according to Cohen and Manion (1994: 136), are appropriate under the following conditions:

- A firm basis of previous knowledge should exist. In other words, there must be grounds for assuming that relationships between the factors and performances being studied exist. In the PTEEP context, much related work has been undertaken, locally and internationally, which establish grounds for such an assumption. Examples of these studies can be found in Alderson and Clapham (1992), Beatty, Greenwood and Linn (1999), McNamara (1996), Waters (1996), Weir (1990, 1988, 1983) and Yeld and Haeck (1997), amongst others.
- There needs to be a reasonable chance that at least a moderately high correlation coefficient will be achieved. This is a difficult condition to meet, as the number of confounding variables in establishing the predictive value of a measure of performance is large, ranging, for example, from personal circumstances affecting study conditions (e.g. financial hardships) to motivation, to area of study. Thus it is unlikely that a high correlation will be achieved, but it is likely (as demonstrated for example in (as demonstrated for example in Dawes, Yeld & Smith 1999, Yeld

& Haeck 1997, and Yeld, Badsha & Shall 1989), in respect of the PTEEP tests, that meaningful, although fairly low, correlations will be achieved.

- The future behaviour being predicted should not be too distant in time. This condition places some pressure on studies investigating the prediction of success in Higher Education, as the time to graduation is at least three years (for educationally disadvantaged students, the length of time is commonly two years beyond the minimum duration of a degree programme). For this reason, many studies rely on first-year performance (e.g. the many investigations of the SAT in the United States), although this performance is a far from satisfactory indicator of eventual success or failure.
- Groups, not individuals, should be the focus of the research. The PTEEP studies use cohorts of students, and investigate how effective various predictor variables are for different groups. As such, they run into many of the difficulties commonly encountered in cohort or longitudinal studies. For example, they are time-consuming and expensive. Furthermore, the difficulty of sample mortality is a continual worry, and cohort studies are only viable if a large enough group is available to begin with. In investigating the validity of the PTEEP tests, data sets were built up from a number of years, as described below.

As noted by many researchers working in the area of admissions (e.g. Dawes, Yeld & Smith 1999, Mitchell, Fridjohn & Haupt 1997, Green 1995, Yeld, Badsha & Shall 1989, Donlon 1984, and Crouse & Trusheim 1988), performance can be assessed in a number of ways. As Cronbach (1988) suggests, the importance accorded to the criterion by predictive validation efforts leads to considerable vulnerability. Level of achievement, ratio of courses taken to courses passed, number of courses taken (i.e. academic 'load'), and so forth, all give different pictures of student progression through their studies. In addition, the time or level at which the performance is assessed is important - certain selection criteria might predict first year performance but not throughput to graduation, for example - a point illustrated in an interesting study by Mitchell et al (1997).

The second group of techniques for assessing predictive validity focuses on the ratings of ability assigned by interlocutors such as peers, teachers and/or employers. It is closely related to concurrent validity techniques, where one common approach is to ask lecturers in whose classes students have been placed, whether the test results of the students accord with their impressions of the students' abilities. However, as Alderson et al (1995:182) note, "[I]n many circumstances where tests are developed, it is impractical if not impossible to gather external data on test candidates". In the case of the PTEEP tests, which are designed to serve the whole university, and not any one particular course, no systematic investigation of this kind has been undertaken. However, the test development team includes both the coordinator of the institution's Language Development Group, which is responsible for designing and implementing a wide range of academic literacy interventions, and the convenor of the English for Academic Purposes courses into which some students are placed on the basis of their performance on the PTEEP tests. In this way feedback could be obtained, and it is recommended that future development of the PTEEP tests develop procedures to obtain such external data.

Some of the problems encountered in establishing predictive validity have been discussed in Chapter Four, section 4.3, and are not rehearsed here. They include those problems which arise as a consequence of difficulties with defining the criterion for success and failure in the criterion situation; truncated samples; small sample sizes; the relatively long time between predictor and criterion; and the confounding impact of educational interventions.

The assessment of predictive validity is further complicated by the fact that by the time predictive validation studies are completed or can be undertaken, the test is usually in use and there are other pressures such as time, expense, and vested interests. Also, the original motivation for the test might have shifted: if the test's new versions reflect this shift, validation of the earlier forms may appear somewhat pointless.

A major problem in establishing predictive validity is in assessing exactly what it is that is being measured. In language tests of the work sample type, language is only a part of what is being measured, and contributes only in part toward success or failure in the criterion situation. Indeed,

many studies of the predictive validity of EAP tests (e.g. Criper & Davies 1988) show that "... language proficiency [as measured by the various tests] was found to account for no more than about 10% of the observed variance in outcome measures" (McNamara 1996:21).

Sections 10.3 and 10.4 below report on the studies which have been undertaken to assess the predictive validity of the PTEEP tests. The assumptions underlying the data are discussed in Chapter One of this study. Briefly, they are that the tests can be considered to be equivalent in many respects, and that the candidates are comparable over the years. In relation to predictive validity, however, a major variable concerns the academic curriculum for which the students in the studies are registered. For example, if half of the students in a sample have been placed on a foundation course, and half have not, it would be necessary to disaggregate the groups in reporting the results in a prediction study: that is, the effect of the educational intervention needs monitoring.

In the studies reported in sections 10.3 and 10.4 below, however, students were registered for a wide range of courses, only some of which contained support components. In addition, they were registered across the full range of Faculties, which require very different levels and kinds of engagement. One has only to note the striking differences in pass rates between, for example, courses in Statistics and courses in Social Work, to realise that trying to match students in terms of academic curriculum would require very large numbers – and even then, it would be difficult to make generalisations. In the correlational studies reported in 10.3.2.1 and 10.3.2.2 below, the difficulties for interpretation consequent on attempting to conduct Faculty-specific analyses are evident.

Because of the difficulty of building comparable samples on the basis of the students' educational experiences at UCT, the data sets were derived on the basis of school background. In support of this decision, the following features of the educational environment at UCT should be noted:

- Many students who are recommended by the Project are admitted onto the regular curriculum. Some of these 'regular admission' students would gain entry through their SC results, as is made clear below in section 10.4, and some on the basis of excellent PTEEP results. In other

words, not all students admitted on the basis of PTEEP scores are placed onto foundation or extended programmes. In addition, not all students who meet the institution's regular cut-off points are placed onto the regular curriculum.

- UCT has no fixed foundation year 'package' in any Faculty. Therefore, even students placed onto a particular foundational-type curriculum would be taking a mixture of extended (credit-bearing but with double contact time) and regular courses. This mixture of courses is seldom the same for more than a very small group of students, as flexibility has been shown to be essential with a highly diverse intake. For example, two students may be admitted onto the extended programme in Commerce, and one of these may 'accelerate' onto the regular curriculum at the end of the first semester, making their curricula then not directly comparable. Another might be exempted from one or more extended courses, and take very close to a regular load. Yet another might be placed onto the regular curriculum but be required to take one support course, such as 'English for Academic Purposes', a one semester credit-bearing course in Humanities.
- Many 'regular curriculum' students, as is the case with many non-AARP students, move, voluntarily or involuntarily, onto a supported curriculum after the first semester or first year.
- The range of demands found across curricula and Faculties, as mentioned above, works against the notion of 'extended' versus 'regular' curricula.

The studies are as follows:

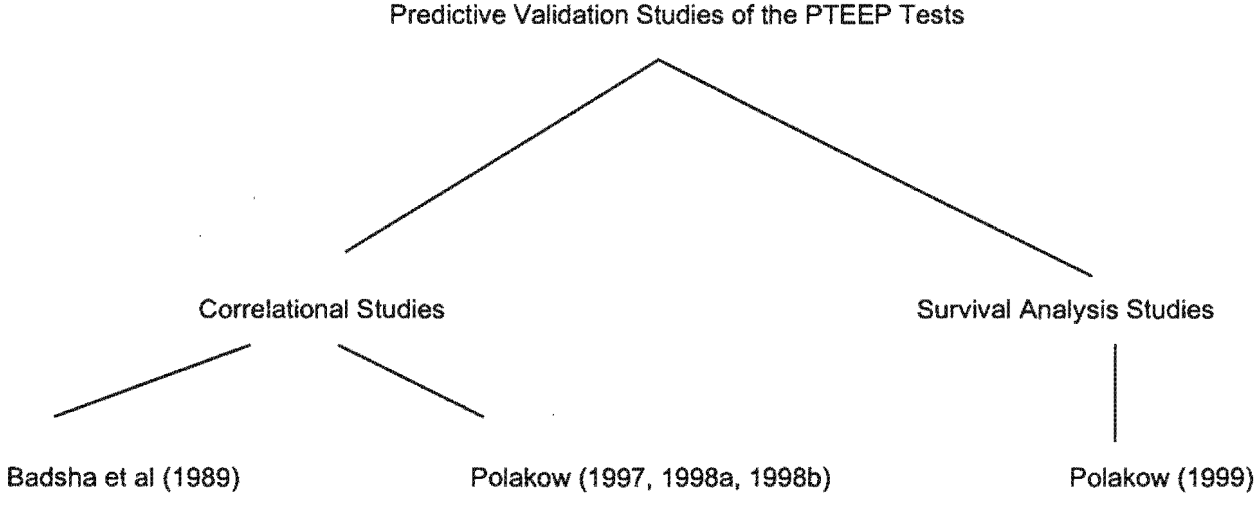

Figure 10.1: PTEEP Validation Studies

In all cases, the studies are based on the performance of students from the former DET education authority.

10.3 Correlational Studies and the PTEEP Tests

In assessing the predictive validity of the PTEEP tests, two sets of correlational, predictive studies were conducted. In both cases, the variables considered included Senior Certificate (SC) results (the aggregate score, comprising the results of six SC subjects, as well as the English Second Language Higher Grade (ESL-HG) result); the PTEEP tests; and performance at university. While neither of the two sets of correlational studies revealed strong relationships between any of the variables being considered, they did point to some interesting trends, as is evident in the discussion below.

It should be emphasised at this point that the aim behind the development of the PTEEP was - and is - to develop additional means (additional to the Senior Certificate) of selecting talented, educationally disadvantaged students for Higher Education study. Therefore, if it transpires that SC results are to some extent predictive in terms of future academic performance, this would not mean that an additional test such as the PTEEP does not have a role to play if it can be shown that the PTEEP is also predictive. On the contrary, if the PTEEP results reveal positive and significant relationships with academic performance, then - irrespective of the predictive power of the SC - the PTEEP will have a contribution to make in ensuring that a maximum of talented students are recruited for the institution (and the system). In fact, the only circumstances in which an economical and predictive test would not be of value in addition to the SC would be either if the SC were perfectly predictive or perchance, if the PTEEP had complete overlap with the SC.

10.3.1 Yeld, Badsha, and Shall (1989) Correlational Studies

The first set of correlational studies is reported in Yeld, Badsha, and Shall (1989). The data used in these studies were derived from: (i) the ELPT (English Language Proficiency Test), an early form of the PTEEP developed by AARP and used until 1989, when the PTEEP tests were

introduced, (ii) the ESL-HG scores, and (iii) SC point scores. These last were represented by the points system as shown below in Table 10.1.

SYMBOL	HIGHER GRADE	STANDARD GRADE
A	8	6
B	7	5
C	6	4
D	5	3
E	4	2
F	3	1

Table 10.1: UCT Points Score System for Admissions

The study was based on the results of 152 ex-DET students, who formed the 1988 cohort. The statistical methods employed in this study are correlation analysis (Spearman's rank correlation coefficient); Kruskal's gamma test for strength of association of ordinal categories; factor analysis; and chi-square analysis.

The findings of these studies were as follows:

- There was a significant relationship between performance on the ELPT and mid-year academic performance in first-year courses. As measured by Kruskal's gamma, the strength of the association was 0.51. In contrast, no significant association for SC points and mid-year academic performance was found. This finding was important, and indicated that the AARP tests had a useful role to play in selecting educationally disadvantaged students.
- There was a significant relationship between performance on the ELPT and overall first-year (i.e. both mid- and end-of-year) academic performance in first-year courses. As measured by Kruskal's gamma, the strength of the association was 0.33. Again, in contrast, no significant association for SC points was found (Kruskal's gamma of 0.18).
- No significant relationship was found between the SC ESL-HG score and ELPT performance. This suggested that the two language tests were perhaps not tapping the same abilities. This area is explored further in Chapter Eleven.

These early findings provided support for the continuation of the Project, and indicated that the tests had a useful role to play in the admission of educationally disadvantaged students, and were not simply replicating the SC results. In addition, they consistently showed that the tests did not

correlate significantly with the ESL-HG examinations. This finding suggests that these two language tests were not testing the same construct, and/or were employing such different approaches to measuring language proficiency that they could not properly be compared. This point will be analysed in some detail in Chapter Eleven in the discussion on concurrent validity.

10.3.2 The Polakow Correlational Studies

The second set of correlational studies was conducted by Polakow (1998a, 1998b, 1997). As was the case with the Yeld et al (1989) study, all the students in the sample had written their Senior Certificate examinations under the authority of the DET, and all had taken English as a Second Language (ESL-HG) in the SC examination. After lengthy consultation with the author and AARP colleagues, an elaborated, more complex approach was adopted to understanding university performance than the Yeld et al (1989) study, and because the various indices of university performance are used in subsequent studies, it is reported here in more detail.

Although some faculties weight individual scores (e.g. the Science Faculty doubles the Mathematics and best Science subject score), it was decided to use un-weighted points (as illustrated in Table 10.1 above) so that cross-Faculty comparisons could be made. For this reason, too, only the SC results of candidates who had written at least six subjects were included: for the few candidates writing more than six subjects, the six best were used. In order to understand university performance data, individual courses needed to be weighted in terms of their contribution to the students' curricula. The weighting of the courses was, in most cases, straightforward (e.g. courses tended to be full year or semester courses) but some courses carried partial credits and needed to be interpreted in consultation with Faculty officers.

Three indices of performance at university were distinguished (Polakow 1997):

- Index 1. This summary value takes into account the actual mark, expressed as a percentage, obtained by a student for each of her/his courses, as well as the weighting of that course in terms of time and degree points. It then expresses these components as a weighted average of the student's performance over all his/her courses. Thus, if a student obtained 62% for a

semester course - i.e. a course with a 0.5 weighting, the student has a 'weighted score' of 31 for that course (0.5 x 62 = 31). This 'weighted score' is then added to all the other weighted scores obtained by that student, and the final value is divided by the number of full year credits taken (i.e. the course weightings).

- Index II. This index examines performance on courses simply in relation to the number of courses a candidate has taken. This device takes into account the load carried by a student, although the quality of pass is not considered. In calculating this index, only credit weightings are considered (so the sum of all the credit weightings of the courses passed by a student is divided by the sum of credit weightings of all courses taken by a student).

- Index III. This index is based on the readmission status of a student. At the University of Cape Town, an elaborate system of readmission codes is employed, which is sensitive to the educational and personal histories of students. Two students with identical results but different educational histories could therefore be coded in quite different ways. In Polakow's study, all the students were classified as ex-DET, i.e. educationally disadvantaged, and their readmission codes were therefore treated as comparable. The readmission codes were grouped as a dichotomous variable (0 or 1), with 0 representing those codes which signify readmission, and 1 those signifying exclusion. However, the fact that students were allocated codes on such an individualistic basis meant that this index was not useful for analytic purposes, and no significant relationships were found for Index III between any of the measures and university performance. The discussion below therefore does not include further mention of Index III.

For indices I and II, supplementary examination results superseded the original result for that subject. For Index II, students who failed to write their examinations - that is, who were not exempted from writing on medical or other grounds but who simply did not write - were treated as having written and failed, as were those who were not allowed to sit the examination on the basis of poor in-course performance. For Index I, however, such students were omitted from the analysis, because no mark was available.

In coding and sorting the Senior Certificate and PTEEP data for analysis of variance manipulations, both SC points and PTEEP scores were sorted and converted into quintiles. For the 1996 cohort analysis, the SC ESL-HG scores were transformed into an approximately even distribution of candidates into roughly equally-sized categories. ESL-HG symbols A and B were combined to form category 1, C became 2, D = 3, and E and F were combined to form category 4. In other words, the ESL-HG scores were sorted and converted into four categories, not five as was the case for the SC and PTEEP scores. For the 1995 cohort analysis, the SC ESL-HG scores were transformed into three levels or categories to create an approximately even distribution.

The statistical methods employed in Polakow's study are correlation analysis (Spearman's rank correlation coefficient); analysis of variance (ANOVA) and Kruskal-Wallis tests; discriminant function analysis and logistic regression; 2-sample non-parametric testing (Kolmogorov-Smirnov); and chi-square analysis.

The three indices of performance described above were analysed in three ways: first, irrespective of the Faculty in which students were registered (pooled data); second, by faculty grouping (un-pooled data); and third, by whether a student's PTEEP score fell into the top quintile, or into the bottom four quintiles. These were called the PTEEP top/bottom groupings, and were essential to investigate as the project had been using, as a rule of thumb, the top quintile as a basis for its recommendations. If this rule were to be shown not to be meaningful, further investigation would be necessary – for example, into whether a top third versus the lower two-thirds top/bottom grouping, or a top decile versus the lower deciles top/bottom grouping.

This procedure was followed for the cohorts of students entering UCT in 1995 and 1996, and both cohorts were tracked over two years.

10.3.2.1 The 1995 Cohort: First and Second Year Performance

Relevant results from analyses conducted on this cohort of students are shown below in Table 10.2 (significant results are displayed in bold). The indices of performance used (Indices I and II) are described above in section 10.3.2. Note that, as was described in Chapter One, section 1.5.2.2, Kruskal-Wallis tests (H) are a non-parametric analogue to

ANOVA techniques (F), and are used with ranks rather than means. Both statistical procedures offer similar interpretations in these analyses.

		INDEX I			INDEX II		
		R	F	N	*R*	H	N
PTEEP	Year 1	**0.19**	not sig.	295	**0.22**	**13.8**	320
	Year 2	not sig.	not sig.	231	not sig.	not sig.	276
SC points score	Year 1	**0.26**	**4.96**	295	**0.20**	**16.32**	320
	Year 2	**0.18**	**2.99**	231	**0.17**	**10.41**	276
ESL-HG	Year 1	**-0.17**	**4.65**	294	**-0.16**	**9.62**	318
	Year 2	not sig.	not sig.	231	not sig.	not sig.	276

Table 10.2: 1995 Cohort, First and Second Year Performance (Pooled Data)[75]

It can be seen from the table that in terms of first-year performance, for both Indices I and II, PTEEP, SC points score and ESL-HG are all significant, albeit weak, predictors of university performance. ESL-HG, however, displays an inverse relationship: (R = - 0.17 in respect of Index I, and R = - 0.16 in respect of Index II). This finding is important, as admissions officers at UCT and elsewhere in South Africa place great store on candidates' school-leaving results in the medium of instruction. The inverse correlation cautions that this reliance may not be a sensible option for students such as the group in this study.

ANOVA test results for first-year performance indicate that the procedures adopted to categorise the measures show significant differences for both SC points score and ESL-HG, but not for the PTEEP tests, in respect of Index I. In respect of Index II, significant differences were observed for all three measures.

In respect of second year performance, it can be seen that several changes have occurred (from first year to second year) in the relationships of the predictor variables to performance. This underscores the danger of relying on only first year data in predictive validation studies. The table reveals that only the SC points score correlates with Indices I (0.18) and II (0.17). This correlation is extremely weak, however.

ANOVA test results for second year performance reveal significant differences in score categories only for SC points score, for both Indices I and II.

[75] Alpha levels of significance were set at 0.05 for all the statistical tests in the Polakow study.

When the data were un-pooled (categorised into faculties), none of the three measures yielded any significant predictors of performance at first or second year level in any faculty except Commerce.

		INDEX I			INDEX II		
		R	F	N	*R*	H	N
PTEEP	Year 1	not sig.	not sig.	49	not sig.	not sig.	49
	Year 2	**0.43**	not sig.	29	not sig.	not sig.	45
SC points score	Year 1	**0.37**	**9.06**	49	not sig.	not sig.	54
	Year 2	**0.56**	**12.95**	29	**0.42**	**10.54**	45

Table 10.3: 1995 Cohort, First and Second Year Performance (Un-Pooled Data) Commerce Students

It can be seen in Table 10.3 that SC points revealed a moderate correlation (R = 0.37) with first year performance in Commerce in terms of Index 1, and a relatively strong correlation of 0.56 at second year level.* A moderately strong relationship between SC points and second year Index II performance (0.42) was also revealed. However, this is difficult to interpret when no significant relationship with Index II was revealed with first year performance. A similar difficulty exists in the relationship of 0.43 between the PTEEP and Index 1 performance at second year level, although no significant relationship can be found at first year level.

ANOVA and/or Kruskal-Wallis tests reveal, in relation to the unpooled data, that the procedures adopted to categorise the measures show significant differences only for the SC points score measure for first and second year performance in respect of Index I, and for second year performance in respect of Index II.

For all faculties except Commerce, then, PTEEP and SC points scores reveal significant although very weak relationships with performance at first year level, as does the SC points score with performance at second year level (see Table 10.2). This, however, was not regarded as helpful in terms of admissions decisions, as no significant relationships with specific Faculties could be found. For Commerce, the SC points score yielded promising information, although as can be seen below, this was not repeated with the 1996 cohort.

Additional findings show that when re-categorised into top/bottom groupings, PTEEP yielded a statistically significant predictor of Index 1 (N = 295, F ratio = 4.82) for first-year performance, but

not for second year performance (except in the Faculty of Humanities, where it was significant). This lends partial support to the AARP practice of recommending the top quintile for admissions. PTEEP and ESL-HG are relatively strongly correlated, although, again, this is an inverse relationship (R = -0.50, N = 318). A similar, albeit weaker, inverse relationship was revealed in the Yeld et al study (1989) described earlier. Reasons for this negative relationship are investigated further in Chapter Eleven.

It can be seen that the results of the 1995 cohort are difficult to interpret in terms of the usefulness of the predictor variables. For most faculties, the fact that a correlation of 0.26 in first-year (and 0.18 in second-year) was found between SC points and Index 1 performance was not persuasive when it was revealed that faculty-specific correlations were not found. For the Faculty of Commerce, however, the SC points score was found to be a positive, although weak, predictor for this cohort.

10.3.2.2 The 1996 Cohort: First-and Second-Year Performance

Once again, a relatively strong, negative correlation (-0.42) was found between results on the PTEEP and results on the ESL-HG. Between PTEEP and SC points score, however, a weak correlation of 0.24 (N = 302) was found. Interesting findings in respect of the 1996 cohort and university performance are displayed in Table 10.4 below.

		INDEX I			INDEX II		
		R	*F*		*R*	*H*	N
PTEEP	Year 1	**0.21**	**3.81**	273	**0.15**	not sig.	302
	Year 2	not sig.	not sig.	206	not sig.	not sig.	253
SC Points Score	Year 1	**0.31**	**8.54**	273	**0.14**	**12.63**	302
	Year 2	**0.27**	**13.65**	206	**0.18**	**25.12**	253
ESL-HG	Year 1	not sig.	**2.68**	273	not sig.	not sig.	302
	Year 2	**-0.14**	Not sig.	193	not sig.	not sig.	253

Table 10.4: 1996 Cohort, First and Second Year Performance (Pooled Data)

It can be seen from the table that, in relation to first-year performance:

- Index I is weakly correlated with performance on PTEEP (0.21) and moderately with the SC points score (0.31).
- Index II is weakly correlated with performance on PTEEP (0.15) and SC points score (0.14).
- There is no statistically significant correlation at the first year level between ESL-HG and Indices I or II.

- ANOVA test results indicate that the procedures adopted to code and sort the PTEEP results, the SC points score and the ESL-HG results have revealed significant differences between the groups within these tests (e.g. the quintiles for SC points and PTEEP scores), in terms of their relationship to first-year university performance as measured by Index 1. The Kruskal-Wallis test indicates that a similar situation exists in relation to Index II in respect only of the SC points score.

The data illustrated in Table 10.4 confirm the problem with using the ESL-HG for selection purposes, and suggest that the three predictor variables are not tapping the same abilities. However, the data also suggest that none of these variables is very useful for admissions or selection purposes.

PTEEP top/bottom groupings appear to be a statistically significant predictor of Index I (F = 9.64) and Index II (H = 7.83).

In relation to second year performance, the table shows that Index 1 is weakly correlated with the SC points score (*R* = *0.*27). ESL-HG is also weakly correlated with Index 1, but, although statistically significant, this correlation is negative (*R* = -0.14). In relation to Index II, the only significant predictor is the SC points score (*R* = 0.18). The relationship is extremely weak, however. The ANOVA and Kruskal-Wallis tests indicate that the procedures adopted to categorise the measures show significant differences in respect of second year performance only in relation to SC points score, for Indices I and II.

In respect of second-year performance, PTEEP top/bottom groupings appear to be a statistically significant predictor of Index I (F = 3.90) but not Index II.

However, when the data were un-pooled (i.e. assigned to one of five university faculty categories rather than being treated as one whole), none of the measures revealed any significant relationships with first-year university performance in any faculty (for any of the indices) except for the Science Faculty, where a correlation of 0.26 (N = 60) was revealed in respect of Index I and

the SC points score, and the Faculty of Engineering, where a correlation of 0.43 (N = 43) was found in respect of SC and Index I, and a negative correlation of 0.31 (N = 42) between ESL-HG and Index 1. However, for neither of these faculties was a significant correlation found in relation to second-year performance and Index I. These results are displayed below.

			INDEX I				INDEX II			
			R	N	F	N	*R*	N	H	N
PTEEP	**Commerce**	Year 1	not sig.	53	not sig.	53	not sig.	53	not sig.	53
		Year 2	not sig.	32	not sig.	32	**0.30**	**45**	not sig.	46
	Engineering	Year 1	not sig.	43	not sig.	53	not sig.	53	not sig.	43
		Year 2	not sig.	27	not sig.	27	not sig.	35	not sig.	35
	Science	Year 1	not sig.	60	not sig.	60	not sig.	60	not sig.	60
		Year 2	not sig.	36	not sig.	36	not sig.	42	not sig.	42
SC Points	**Commerce**	Year 1	not sig.	53	not sig.	53	not sig.	53	not sig.	53
		Year 2	**0.48**	**32**	**15.28**	**32**	**0.53**	**46**	**18.35**	**46**
	Engineering	Year 1	**0.43**	**43**	not sig.	53	not sig.	53	not sig.	43
		Year 2	not sig.	27	**10.29**	**27**	**0.51**	**35**	**10.91**	**35**
	Science	Year 1	**0.26**	**60**	not sig.	60	not sig.	60	not sig.	60
		Year 2	not sig.	36	not sig.	36	not sig.	42	not sig.	42
ESL-HG	**Commerce**	Year 1	not sig.	50	not sig.	50	not sig.	50	not sig.	50
		Year 2	not sig.	32	not sig.	32	**-0.33**	**44**	not sig.	46
	Engineering	Year 1	**-0.31**	**42**	not sig.	50	not sig.	50	not sig.	42
		Year 2	not sig.	27	not sig.	27	not sig.	35	not sig.	35
	Science	Year 1	not sig.	55	not sig.	55	not sig.	55	not sig.	55
		Year 2	not sig.	32	not sig.	32	not sig.	38	not sig.	38

Table 10.5: 1996 Cohort, First and Second Year Performance (Un-Pooled Data)
Commerce, Engineering and Science Faculties

In relation to second year performance when the data were un-pooled, only the SC points score, and then only for the Faculties of Commerce (Indices I and II) and Engineering (Index I), showed any significant positive relationships with university performance. In the case of Commerce, a negative correlation of 0.33 with ESL-HG in relation to Index II ESL-HG was found. These findings are illustrated in Table 10.5 above. However, since these relationships were not revealed in relation to first year performance in these faculties, the value of the second year relationships is difficult to determine.

These correlational studies yielded important and interesting information. First and foremost, however, they reveal the extent and complexity of the selection challenge. It can be seen that, on the whole, the correlational studies failed, with the partial exception of the Faculty of Commerce for the 1995 cohort, to provide the kind of unambiguous answers for which admissions officers and Faculty Deans were asking. In essence, the most pressing question was this: How confident can

admissions officers be that students recommended by the Project[76], and/or on the basis of their SC performance, will graduate? While there was, and remains, a great deal of concern about the reliability and effectiveness of the SC points score system particularly in relation to educationally disadvantaged students, custom and national acceptance of the system assured its continued use. The new, alternative method offered by the Admissions Project needed, therefore, not only to answer the primary question posed above, but needed to contextualise it within the existing, prevailing system. In other words, what the Deans and admissions officers really needed to know was how confident they could be about various forms of admissions information. They were not necessarily concerned about the competing claims of these different forms, although these were of course important to report where possible.

Clearly, the correlational studies reported on above did not yield information that adequately addressed these questions. Where predictive relationships did exist, they were on the whole weak, and explained very little of the variance in performance. In addition, they revealed little about overall progress through a degree, but focused on particular academic years as the overriding unit of investigation. After extensive discussions between the author of this study and Daniel Polakow, part-time statistician for the Project, on the precise nature of the information that was needed, Polakow proposed a different and novel approach to understanding and reporting student performance in light of the various hypothesised predictors (PTEEP, SC points scores, and the ESL-HG examination).

10.4 Survival-Analysis Studies Undertaken to Assess the Predictive Validity of the PTEEP Tests

Collins and Horn (1991: 311) suggest that "... research questions framed as questions of 'success' – success in school, in work, or in aging – can be seen to be, implicitly, questions of how many and how long, for some people never get to a specified end point". Such questions focus, as Willett and Singer (1991:312) assert, on "... variation in duration as a function of predictors ...": Some of the predictors they mention are group membership, treatment, environment, or background.

[76] That is, students whose results fall into the top quintile (top two deciles) of the AARP scores for their respective educational grouping.

Survival analysis techniques are based on the survival time of a group of subjects, and were primarily developed in the medical and biological sciences. Survival time is usually defined as the length of the interval between diagnosis/treatment and death: in admissions terms, this notion can be adapted to refer to the interval between selection/admission and exclusion (failure) from an institution. One of the major complications, as Lawless (1982) points out, however, is that one is not customarily able to wait until every subject in one's study has either graduated or been excluded – one has to analyse the data while some subjects are still continuing their studies. In Higher Education, too, the relatively high dropout rate of otherwise academically successful students[77] means that, like patients who are lost to follow-up as they have moved out of the area, there is uncertainty about the rate at which they would have been excluded or have graduated had they remained at the institution. One would not want to exclude all of these subjects from the study by declaring them to be missing data, as they are 'survivors' who demonstrate the success of the selection process. Survival analysis calls the data acquired from such instances 'censored observations': the term reflects the fact that the data contains only partial information (e.g. student A was successful for two years before she left the university in good academic standing). Survival analysis techniques enable researchers to include and deal with censored as well as uncensored observations in a single analysis. The method is illustrated below.

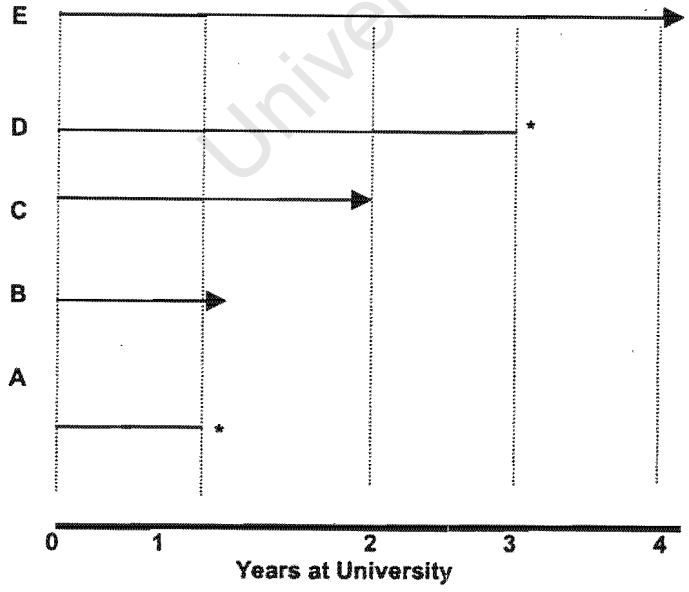

Figure 10.2 : The Tenure Process (Polakow 1999)

[77] No systematic study has yet been undertaken about the reasons for such students dropping out. Hypotheses include such matters as financial problems, illness, registering at another institution, and family pressures.

A student's university tenure is made up of the number of years that the student has 'survived' – that is, has not been excluded. Excluded students are treated as 'failed' observations, indicated in Figure 10.2 by the symbol * as can be seen in individuals A and D, who have been excluded at the end of their first and third years respectively. Individual C above represents a student who has left the university in good academic standing (a censored observation). Her/his progress is represented by an arrow to indicate that s/he could have continued – i.e. an arrow indicates success for a period of time, but without exclusion being observed. It can be seen that since the observation process for any student begins on admission, the most recent cohorts contribute a shorter period of data.

Polakow (1999) conducted a "survival analysis of university tenure" study to assess the predictive validity of the PTEEP tests, which focused on contrasting the probability of exclusion of two groups. This new approach addressed the following primary question:

How do the observed exclusion rates of one group of students over their university tenures compare with the exclusion rates of a different group or groups of students?

In Polakow's study, several groups of students were identified. All the students were registered at UCT in the 1995 – 1998 inclusive cohorts, and all were ex-DET students. The groups were constituted as follows:

- students who had not written the PTEEP (n = 563). These students gained entry to UCT on the basis of their SC points scores.
- students who had written the PTEEP (n = 511) before entering the university. These students gained admission to UCT on the basis of their SC points scores and/or their performance on a PTEEP test. Thus, not all of this group would have been recommended by AARP for admission. This group was divided into two as follows:

- students who had written the PTEEP and had achieved scores in the top quintile (n = 120)[78]. All of these students would have been recommended by the Project for admission. Some of these students would also have obtained the required SC points for admission.
- students who had written the PTEEP and had achieved scores in the bottom quintile (n = 40). None of these students would have been recommended for admission by AARP – therefore, all of them would have obtained the required SC points for admission. This occurs because the AARP tests are not used to exclude students.

The groups are illustrated below:

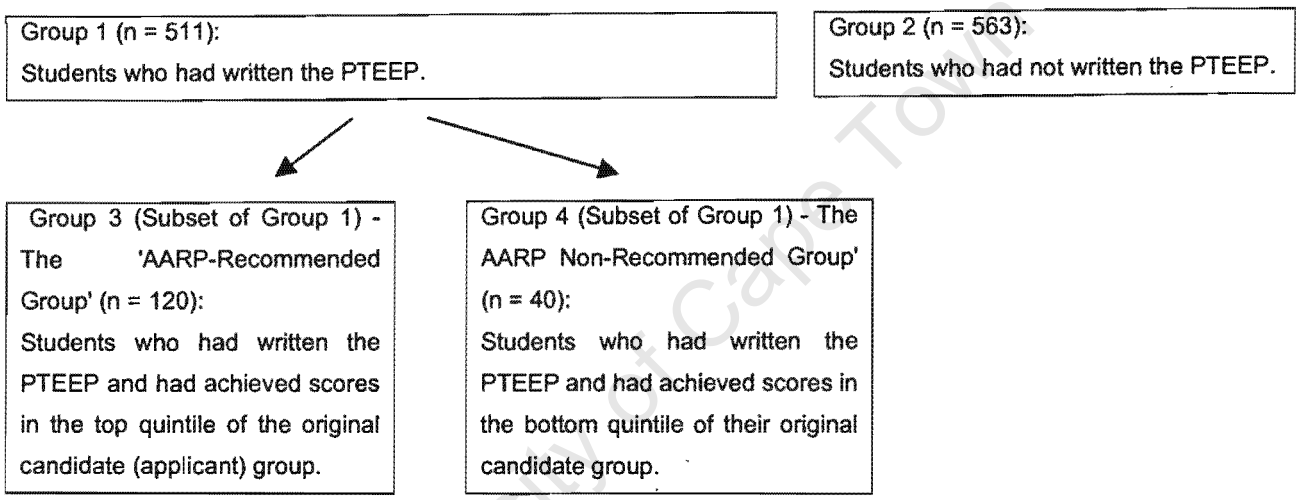

Figure 10.3: Survival Analysis Data Groups

All the students in the study were ex-DET students. Reasons for this focus include the following: it is this group of students who tend to achieve results in the lower SC points ranges; prior to 1996, the Project only targeted students from DET schools, and so longitudinal data were not available on candidates from other educational departments; and restricting the range of prior educational opportunities constrained the impact of this crucial variable.

All four cohorts of students (1995 - 1998) were pooled into the analysis. Thus, the 1995 cohort has the most impact on the study as it contributed data over all four years. The 1998 cohort, however, contributes data to only one year.

[78] It should be noted that the quintile boundaries were derived from the performance of the total applicant pool (usually over 2,000) of ex-DET candidates writing the AARP tests, and not from the performance of the smaller group of students entering the university.

An important finding during the assembly of the data-sets was that the SC points score of the students who wrote the AARP tests and entered the university (recommended and non-recommended) is significantly different from that of the students who did not write the AARP tests. According to Polakow (1999), this difference is significant for their medians (Median-test, χ^2 = 4.235m d.f. = 1, P = 0.040), and for the distribution of their respective ranks (Kruskal-Wallis, N = 1056, H = 10.690, P = 0.001). The means are significantly different (two-sample t-test, d.f. = 1054, t = -2.844, P = 0.005) – the large sample sizes allow the tests to give adequate evidence of even small differences between means. The results show that the non-AARP test writers have slightly but significantly higher SC results than the AARP-writing group. This finding is important, as it suggests that if stronger academic performance (at UCT) is found for the AARP-writing group, it cannot simply be ascribed to their being selected from a stronger group in the first place.

Success and failure were categorised as follows: 'successful' students were those who had graduated or were continuing their studies, whereas 'unsuccessful' students were those who had been excluded. Students who had left the institution in good academic standing (for example, to earn money to pay for further years of study, or for family reasons) were not eliminated from the analysis. As censored observations, as described above, they were "... assimilated as an incomplete retention time ..." (Polakow 1999:4) – that is, their length of tenure contributed to the data, where they were classified as successful.

The findings of the survival analysis study are as follows.

Generally, hazard rates (the likelihood of being excluded) increase over the first and second years of university, and then decrease in subsequent years. More specifically, in all the analyses, hazard rates were higher in the second years of study than in the first. This feature is of interest, particularly in relation to predictive studies that use first year performance as the criterion of success and failure. There are many possible reasons for higher failure rates in second as opposed to first year. For example, many educationally disadvantaged students are placed onto courses with some degree of support in their first year, and find the transition to second year a

difficult one (Griesel 1999). Also, the increased complexity of second year work places higher demands on all students. Whatever the reasons, however, the cautionary note sounded by this study regarding reliance on first-year performance as an indicator of subsequent academic success is clear.

In terms of the question: *How do the observed exclusion rates of one group of students over their university tenures compare with the exclusion rates of a different group or groups of students?*, however, the study showed that for the group of students under the spotlight (educationally disadvantaged students recommended by the Project, irrespective of their SC results), exclusion rates over their university tenures are lower than those of students admitted on the basis of their SC scores. More precisely: for approximately 60% of the time, the top quintile of PTEEP performers have lower exclusion rates than the group that entered the institution on the strength only of their SC results.

On its own, this finding – that for 60% of the time the PTEEP recommended group performs more strongly than the non-AARP group – provides affirmation for the value of the PTEEP scores as predictors of success. From the point of view of admissions officers, it suggests that for more than half the time, they would be better off making selections on the basis of the PTEEP results than they would on their customary grounds (SC results).

In addition to this finding, however, is the following; for approximately 40% of the time, the rates of exclusion are the same for both groups - that is, for the top quintile of PTEEP performers and the non-AARP group that had been admitted to the university on their SC results alone. In other words, at no time did the top quintile of PTEEP performers prove to be a higher risk than those routinely admitted to the university on the basis of their SC results.

This is powerful evidence. From an admissions officer's point of view, it suggests that, in the case of educationally disadvantaged applicants, they can be confident about selecting on the basis of PTEEP scores. It also suggests that SC results for this group of students should be used with caution.

A controversial issue related to the Admissions Project is the potential use of its tests to reject applicants who might otherwise have been admitted to the institution. In order to shed light on this, Polakow investigated the university performance of the bottom quintile of PTEEP writers. He found that for more than 90% of the time, the bottom quintile of the PTEEP performers have higher rates of exclusion than the group admitted on the basis of their SC results. For less than 10% of the time, there is a high degree of unpredictability for this group. This pattern suggests that the tests could perhaps be used to reject applicants, but in view of the seriousness of the step, it was recommended that further research be undertaken before this strategy is adopted.

From the point of view of predictive validity, the findings above suggest that the PTEEP tests are effective in terms of predicting academic success at the University of Cape Town, using non-exclusion as the criterion of success.

10.5 Conclusion

The analyses in this chapter have highlighted many of the difficulties encountered in predictive validity studies. These difficulties are exacerbated when the criterion situation is difficult to define. In the studies reported here, it can be seen that narrowing the criterion definition (for example, focusing on the performance of Commerce Faculty students, as opposed to overall, university-wide performance), has the consequence of reducing the number of candidates and thereby introducing other problems. However, in Faculties such as Humanities, where numbers are relatively large, the great heterogeneity - in terms of such factors as levels of difficulty, class size, and assessment types, of curicula means that students cannot sensibly be treated as a coherent cohort. For example, some curricula in Humanities have more in common with curricula in Commerce than they do with other Humanities programmes.

The difficulties encountered in predictive validity studies are also exacerbated by other problems in defining the construct of success, where decisions have to be made about time/length, level of pass, academic course load, and so forth. This chapter has reported on a promising statistical

approach to predictive validation studies: namely, the technique of survival analysis, which promises to overcome many of the problems encountered in correlational studies.

Despite these and other problems (such as lack of relevant data), however, the analysis of the predictive validity of the PTEEP tests reveals several encouraging findings. For educationally disadvantaged students, the tests appear to be a more useful indicator of academic progress (defined as non-exclusion) than do Senior Certificate results. This is not to say that SC results are not useful predictors, but that for educationally disadvantaged students, who tend to achieve SC results in the lower ranges, the PTEEP tests can provide important and different, additional information. The usefulness of the tests for other groups of students has yet to be established, and it is recommended that such studies be undertaken as a matter of urgency.

The analysis also reveals that the SC ESL-HG examination, on which considerable reliance is placed as a selection measure, is not necessarily a valid measure to use on its own, with educationally disadvantaged students, for this purpose: indeed, at times it displays an inverse relationship with various measures of academic performance. This finding is explored further in Chapter Eleven, where the concurrent and consequential validity of the PTEEP tests is investigated.

CHAPTER ELEVEN

CONCURRENT AND CONSEQUENTIAL VALIDITY

AND

THE PLACEMENT TEST IN ENGLISH FOR EDUCATIONAL PURPOSES

11.1 Introduction

In Chapter Ten, the crucial area of predictive validity – one aspect of external validity - was analysed. Chapter Eleven discusses two further aspects of external validity of the PTEEP tests: concurrent and consequential validity. In the first of these, concurrent validity, a brief comparative analysis is made of the ESL-HG examination and the PTEEP. This comparison is undertaken for two main reasons: first, the ESL-HG examination is the only comparable test that will have been written by all the PTEEP candidates in this study, and therefore, however imperfect, provides the only available data for an investigation of concurrent validity. Second, the analysis is undertaken to investigate why the two tests are so different in terms of their predictive power, as was shown in the preceding chapter. While the ESL-HG examination is not the focus of this study, the weight accorded it by admissions officers, and the importance of English in South Africa and therefore of English Second Language, means that it is essential for these differences to be examined. The analysis focuses on the constructs of the two tests, as well as on how they are operationalised in the form of examination papers.

Consequential validity, the second aspect of validity discussed in this chapter, is a relatively new form of validity, and has arisen alongside generally increased societal concerns regarding transparency and accountability, and the public's right to information. As the PTEEP was originally introduced in an effort to increase black participation at UCT, it is clearly essential to examine the extent to which it has fulfilled this aim, as well as to monitor other intended and/or unintended consequences.

11.2 Concurrent validity

Concurrent validity refers to the relationship between a candidate's score on one test with her/his score on some other measure taken at approximately the same time. This other measure need not be a test, but should involve information that can be collected in a systematic fashion and expressed numerically. For example, a rating scale could be developed to gather self-assessments from candidates, or assessments from academic staff or other sources.

What is essential is that this other source of information be both legitimate and relevant to the aims of the new test. Of course, this requirement raises immediate problems – if there is a legitimate and relevant measure already available, why is it necessary to develop a new test at all? As Alderson et al (1995) point out, some good tests are too long or expensive to be used in all contexts, and so a new, shorter, less costly version is needed. In addition, examinations do not have an indefinite shelf-life – security concerns, new format possibilities, new research questions, new technologies - all make new versions of tests inevitable. To this list could be added the possibility that certain existing tests, while providing useful results, contain inappropriate or insensitive material which is not considered acceptable by the test user community.

However, even if there are no existing tests that are considered valid for the purposes for which the new test is being developed, it might yet be necessary to consider the relationship between existing test/s, however unsatisfactory, and the new test. There are many reasons for this view. For example, public belief in an existing test, often simply as a result of years of use, could be so

strong that comparisons are inevitable, and systematic and reliable information is essential in order to justify or defend the new test.

In some contexts, tests are used in inappropriate ways, despite the careful disclaimers and warnings of the test developers. In this case it is important that a new test be able to demonstrate its credentials in relation to the old, inappropriate test, as it is the old test that will be viewed as the exemplar. This demonstration is particularly important if the new test is significantly different from the existing test. For instance, it might require candidates to produce language by performing authentic tasks rather than by selecting options in multiple-choice formats. In such a situation it will be necessary to show that the new test nevertheless produces similar results. If the new test produces different results, on the other hand, it will be important to acknowledge and justify the value of these results.

The techniques that are customarily used for assessing concurrent validity involve a comparison between candidates' scores on the new test and some other test. One of the difficulties in this regard is that for comparisons to be meaningful, both the tests need to be written – by the same candidates - at roughly the same time (i.e. they need to be concurrent), and under similar conditions. Time and resource constraints make these requirements difficult to achieve, and ethical considerations can make it undesirable to mislead candidates into believing that both tests are important if they are not. Nevertheless, ideally, candidates should write a previously validated test alongside a new test.

Generally, a high correlation (in the order of 0.9) is desirable if a new version of a test is being developed, where results on one are meant to approximate results on the other. In this case the new version could be said to display a high degree of convergent validity. However, if a new approach is being introduced, for example as a result of dissatisfaction with the results of a previous test, a lower correlation than 0.9 might not be a problem. In this latter case, the test construct (the theory on which the test is built) is crucial for interpretation of the new scores.

Other methods used in establishing concurrent validation involve comparing scores on the new test with the rankings or ratings of the candidates' tutors or lecturers. In situations where tests are used to place candidates onto specific courses, and where there are appropriate staff:student ratios for staff to develop meaningful and reliable understandings of the relevant abilities of their students, these methods are realistic and helpful. Candidates' self-assessments are another source of comparison with their scores – however, careful training is needed to ensure that candidates understand fully what the test was attempting to tap.

As is suggested above, concurrent validity is usually established through making a comparison between scores on a new test and scores on a previous or already existing test. In the PTEEP context, where candidates are writing the test for admissions and where they come from far-flung areas to testing centres to write the test, it is clearly not possible to require them to write another test, such as a previous version of the PTEEP. The only other comparable test which almost all PTEEP-writers will have written, and for whom data will be available, is the Senior Certificate (SC) English examination[79]. For students in this study, the English examination will have been the ESL-HG examination, which, as is shown below, is based on a broadly comparable construct.

In the context of this study, the two tests have very different outcomes. These differences in outcomes could be ascribed to two main causes: to differences in the construct of the tests; and/or to differences in the ways they are examined. It should be in addition noted that differences in the ways in which candidates are prepared for examinations is undoubtedly a significant potential source of variance. Investigating this source in the case of the PTEEP versus ESL-HG is confounded by a number of factors, however. First, there is no formal preparation for the PTEEP; second, the design of the PTEEP itself tries to incorporate preparation via the scaffolding approach; third, all the candidates writing the PTEEP in this study have attended DET schools and have all written the ESL-HG examination. The main causes for differences in predictive power are

[79] It should be noted that the fact that the ESL-HG examination is set and administered on a provincial basis, and thus is really not one examination, but nine, is not factored into the analysis. This is justified on the grounds that Higher Education institutions do not distinguish, in allocating SC points scores, between subject results from different provinces; all of the ESL-HG papers are based on the same syllabus; and all are moderated by SAFCERT.

more realistically to be found in the purposes of the examinations, which are closely aligned to the constructs, and to the actual operationalisation of these constructs in the examinations themselves.

Having said this, however, it must be acknowledged that despite the broad similarities in the constructs of the two tests (cf. section 11.3.1 below), the ESL-HG examination is different in significant respects from the PTEEP tests. An example of such differences can be seen in contrasting the purposes of the two tests. In the case of the PTEEP, the test's purpose is to identify potentially successful students for the University of Cape Town. It is thus legitimate to measure its performance in these terms. The ESL-HG examination, however, has a different, larger purpose. It is designed to assess learners' "... communicative competence for personal, social, educational and occupational purposes ..." (DoE 1985:2). It could thus be queried whether comparisons between the two tests are appropriate.

In addition, the ESL-HG syllabus on which the examination is based includes the study of literature, as well as a significant oral component, both of which are examined, and which are absent from the PTEEP language knowledge specifications. The study of literature in particular introduces a 'content' dimension, where learners are required to learn the plots of novels and plays, as well as to understand such literary devices as alliteration, metaphor, simile. This content dimension is simulated in the PTEEP through the provision of rich thematic content, but cannot be viewed as directly comparable.

However, since both tests are used in the admissions process by most South African institutions (indeed, in many instances the ESL-HG result is doubled for admissions purposes), it is argued here that it is legitimate to explore their suitability for such a purpose, while bearing in mind that the ESL-HG examination might be succeeding, or not as the case might be, in some of its other, wider purposes.

11.3 The PTEEP Tests and the ESL-HG Examination

The PTEEP analysis in previous chapters was based on data from the performance of ex-DET students, which revealed the lack of relationship between the PTEEP and the DET ESL-HG results

for the samples of students on which these studies were based. This section examines possible reasons for this lack of correlation between the two tests. The investigation is based on a comparative analysis of the constructs of the two tests as well as of the ways in which they are assessed.

The investigation commences with an exploration of the construct on which the ESL-HG examinations are based. Unlike the PTEEP, which is a proficiency examination, the ESL-HG examination is an achievement measure. That is, it is based on a course of instruction that precedes it, and thus the construct is derived from the syllabus.

11.3.1 Similarities and Differences in Construct

The ESL-HG examination, like the examinations of the other subjects in the Senior Certificate set, is an achievement examination. It is therefore the syllabus which defines the construct on which the examination is based. The ESL-HG examinations relevant to the discussion in this study are based on two syllabuses, the Core Syllabus, and a later, much improved version, the Interim Core Syllabus. Both the Core Syllabus (CS) used from 1986 until 1994, and the Interim Core Syllabus (ICS) introduced in 1995, are based on a communicative approach to language teaching. There are, however, significant differences between the two documents, which reflect some of the developments taking place in the New Literacy Studies area as well as in theories of learning and knowing more generally. As the ESL-HG examinations considered here were based on both syllabuses (students writing the PTEEP tests until 1996 would have studied the CS - the syllabus covers the last three years of schooling, and those writing thereafter the ICS), it is important to consider similarities and differences.

The major differences between the two documents are as follows:

- The ICS contextualises the subject of ESL in South Africa by listing the various ways in which it occupies a unique place amongst languages (it is a second language but also the medium of instruction for the majority of learners, it is the country's only widely spoken international

language, it is still a powerful language in commerce and industry, etc.). This contextualisation was not present in the CS.

- Unlike the CS, the ICS views the first language of the learners as a resource, to be harnessed wherever possible in ESL learning, rather than as, on the whole, a possible source of error.
- Closely related to this last point, whereas the CS saw the 'native speaker' as the ideal to which ESL learners should aspire, the ICS notes that "... the multilingual nature of South African society has led to variation in English vocabulary, syntax, accent, stress and intonation patterns ..." (DoE 1995:4) which should be acknowledged in their own right, provided that communication is not seriously impeded. This contrast reflects a major shift.
- The ICS contains a strong emphasis on heuristic uses of language – that is, the ways in which language is used for learning. Writing, for example is described as "... enabling pupils to clarify and structure their own thinking ..." (op cit: 6). Connections between language and cognitive development are explicitly made, and there are frequent references to language across the curriculum (LAC) approaches. This emphasis is in strong contrast to the CS.
- Writing is seen in the ICS as a process, and the importance of this and of 'process writing' – where language is used to describe a process or change of state – is stressed.
- The notion of reading 'critically' appears in the ICS, alongside a generally more cognitive and integrated view of reading in the curriculum than was evident in the CS.

It can be seen that the ICS reflects a broader notion of communicative language ability than its predecessor - one, moreover, far more in line with ESL's particular position as both vehicle (medium) and target of learning. This view is not to imply that the CS was an inadequate document for its time – on the contrary, it represented mainstream Second Language theory when it was introduced. Indeed, were English not also the medium of instruction, its limitations would not be so glaringly obvious. As was argued in Chapter Six in particular, however, one of the major limitations of the communicative language teaching movement was its privileging of BICS over CALP skills (Cummins 2000, 1984, 1980, Cummins and Swain1986). The relative absence, in the CS, of acknowledgement of the importance of the use of language as a tool for learning is a case in point. This absence is corrected in the ICS, which states in this regard:

- "English is of central importance to the whole learning process" (DoE 1995: 2).
- "The development of language and thinking skills is inextricably linked. It has been postulated that it is through the use of language that children take control of their thinking and create their own universe of understanding. Language ... has a fundamental role to play in the whole process of cognitive development" (op cit: 4); and
- "...teaching and learning of the language should contribute towards enabling pupils to use it for effective communication in a variety of contexts and for a variety of purposes .... for practical purposes as well as for their own personal, education, social and imaginative and aesthetic development" (op cit: 3).

With regard to writing, the ICS states that:

- "[Writing] ... contributes to enabling pupils to clarify and structure their own thinking and enables them to communicate with a wider audience than the one with which they are in daily contact" (op cit: 6); and
- One of the stated aims of the writing component is to enable learners to "... express themselves comfortably in ... forms of writing required by the needs of other content areas" (op cit: 7).

In respect of speaking, the ICS holds that:

- "... organising content effectively and logically", and "conveying meaning" are important aims; and
- An important function of speaking is "... enhancing their own knowledge and understanding of a subject by: asking questions/enquiring; rephrasing statements and questions for clarification; offering explanations or alternatives".

With regard to reading, the ICS states that:

- Reading and the study of literature "... should not be seen as discrete activities in themselves. Rather, the act of reading should contribute to pupils' overall communicative ability" (op cit: 6) and that a wide range of genres should be provided; and

- The aim of reading in the syllabus is to produce critical readers, able to recognise and respond to various techniques of persuasion and manipulation in text.

The ICS has not escaped criticism, however. Kapp (2000b:28), for example, concludes that it "... expresses the need for language-across the curriculum, but the use of language for the development of cognitive academic language proficiency is not reflected in the contexts of use it describes". Kapp's assertion is made on the basis of the prescribed writing activities listed in the syllabus, which require students to "apply the conventions appropriate to practical or functional writing relevant to their daily needs and the demands of the work place" in such genres as telegrams, notices, minutes and reports. This conclusion can be questioned, however. First, other writing activities are listed in the same section of the syllabus, and they include more cognitively oriented activities such as "recording, note-taking, describing a process and other forms of writing required by the needs of other content areas". In the Senior Certificate, and included in the ICS, are such activities as using language in "... sustained, discursive writing as required in a given context for a specific purpose and audience", with due attention to such matters as "basic methods of developing the argument, other persuasive (rhetorical) techniques".

Second, the syllabus uses the term 'conventions' in its literal sense – that is, the activities listed are the ones most distinguished and identified by their conventions – their highly stylised forms. As the quotation invoked by Kapp is the third of four listed sets of writing activities, it seems somewhat selective to choose it alone as representing the syllabus's view of the whole gamut of writing activities (the other three cover different areas).

As Kapp's research illustrates, however, the strength and value of a syllabus resides only in part on what is written in its documents and guidelines. What is important is what happens in the classroom – i.e. how the syllabus is operationalised in teaching and learning terms. This outcome means that any activities that are suggested in syllabus documents must be representative of the syllabus as a whole, and Kapp is correct to draw attention to the relative silence of the documents in relation to activities relating to heuristic and ideational uses of language in particular.

The growing body of classroom research located in South African schools provides rich data attesting to the ways in which the aims of the curriculum are subverted (e.g. Kapp 2000a&b, Plüddemann et al 1999, Taylor & Vinjevold 1999, Macdonald 1990). It is neither appropriate nor feasible to give a full account of these studies here. Some of the main points, however, cover such widespread practices as:

- 'scoping', where teachers inform learners about what will be required in the examination – i.e. how much of the syllabus they need to cover, and precisely how this coverage should be achieved (Kapp 2000a&b);
- the serious lack of written work in all languages (Kapp 2000a, Taylor & Vinjevold 1999, DoE 1998); and
- pronounced emphasis on reading 'on the line' – that is, on "literal, factual reading or information retrieval" (Kapp 2000b:24, Taylor & Vinjevold 1999), as opposed to inferential reading,or reading 'between the lines'.

In Table 11.1 below, the ESL-HG syllabus is sketched in relation to both the Bachman and Palmer (1996) model of communicative language ability, and to the PTEEP language knowledge specifications[80]. The aim of this table is to illustrate the coverage intended by the ESL-HG syllabus designers, rather than to list all the syllabus contents.

[80] The Bachman and Palmer model, and the PTEEP language knowledge specifications, are discussed extensively in Chapters Six and Nine.

Language Knowledge (Bachman & Palmer 1996)		PTEEP Language Knowledge Specifications (Yeld et al 1997)	ESL-HG Syllabus (CS and ICS)
ORGANISATIONAL KNOWLEDGE	Grammatical • Vocabulary • Morphology • Syntax	Vocabulary: 'unknown' vocabulary (deriving meanings from context); 'known' vocabulary Syntax: understanding the syntactical basis of the language	Using reference works (e.g. dictionaries) to find appropriate meanings in context Knowledge about how words are formed, and ability to apply this knowledge Discriminating between workds which sound similar; using punctuation to clarify meaning Using syntax to support meaning – e.g. tense, voice, concord and word order, word classes.
	Textual Cohesion Rhetorical organisation	Understanding relations between parts of text through devices of cohesion such as pronoun reference, particularly demonstratives, referring to statements/propositions or 'entities'); skimming and scanning (e.g. using macro features of text such as headings, illustrations) to get gist of passage, locating particular pieces of information Recognising and using indicators in discourse, especially for introducing, developing, transition and conclusion of ideas, and signalling relations between phenomena	Understanding and using layout, title and contents pages, indices etc. i.e. macro features of text; skimming of text. Using language to organise content effectively and logically; responding to features of text which indicate, *inter alia*, introduction or development of ideas, transitions between ideas, explanations and emphases, the drawing of conclusions.
PRAGMATIC KNOWLEDGE	Functional knowledge • Ideational • Manipulative • Heuristic • Imaginative	Separating the essential from non –essential (e.g. main idea from supporting detail, statement from example, fact from opinion, proposition from its argument, classifying and categorising). Extrapolation and application (e.g. drawing conclusions/applying insights derived from texts, seeing trends); inferencing: (understanding ideas/information in a text, implied but not explicitly stated). Detailed reading for meaning, at sentence level and at discourse level Understanding the communicative function of sentences with or without explicit indicators, such as definition, exemplification, exhortation, argument/persuasion Understanding the importance of 'own voice' (including 'ownership' of ideas) and/or creativity of thought and expression Knowledge of visually encoded forms of information representation (graphs, tables, diagrams, maps, flow-charts) Understanding basic numerical concepts expressed in text/numerical manipulations (comparisons, e.g. greater than, smaller than, percentages, basic fractions (e.g. half of, more than double), basic chronological references, sequencing, basic computations	Using language to convey meaning (e.g. to offer explanations or alternatives, to rephrase or seek clarification). Processing and organising information, distinguishing main points from supporting arguments, statements from examples. Inferring meaning expressed through implication and figurative language. Scanning texts to extract information on a particular topic. Understanding features in text which indicate explanation, description, argument, illustration of a point, drawing of conclusions, etc. Using language for practical purposes as well as for personal, educational, social and imaginative and aesthetic development. Interpreting diagrams and flow-charts.
	Sociolinguistic knowledge (sensitivity to dialect, language variety; register; naturalness criteria); familiarity with cultural references and figures of speech	Understanding metaphorical expression Understanding text genre (including audience, purpose etc.)	Understanding and distinguishing between literal and figurative language. Recognising differences created by style and organisation of various text types, (e.g. textbooks, newspaper articles, application forms, poetry).

Table 11.1: The Constructs of the ESL-HG and PTEEP Tests, and the Bachman and Palmer (1996) Language Knowledge Model

As can be seen, in these terms the ESL-HG syllabus achieves relatively comprehensive coverage in all areas of the model. It therefore seems unlikely that the lack of positive correlation between the ELS-HG (whether based on the CS or the ICS) and the PTEEP tests can be put down to inadequate construct modelling, or very different constructs. A further avenue of enquiry is then obviously the operationalisation of the syllabus in the examination.

11.3.2 Similarities and Differences in Assessment Procedures

In attempting to establish concurrent validity for the PTEEP, a comparative study of performance on the PTEEP and performance on the ESL-HG examination was undertaken. The study is discussed in Chapter Eleven. Relevant results are illustrated in Table 11.2 below ($p < 0.05$).

AARP Test	N	ESL-HG
1988 (ELPT)	152	not sig.
1995 (PTEEP)	318	- 0.50
1996 (PTEEP)	283	- 0.42

Table 11.2: PTEEP/ELPT and the ESL-HG Examination: Spearman's *R*

It can be seen from this table that for the 1988 cohort, no significant correlation was found between the results on the two tests, irrespective of whether they were based on the CS or ICS. In 1995 and 1996, a significant negative relationship was found, indicating that the two tests were inversely related: that is, poor performance on one test was associated with strong performance on the other. In exploring reasons for this negative relationship, Hansen's 1997 study of the ways in which ESL-HG was examined during these years, and of the ways in which the examinations related to the ESL-HG syllabuses, was extensively drawn upon.

Hansen (1997) conducted a comparative documentary analysis of eighteen ESL-HG Language, Writing and Literature papers, and the syllabi on which they were based, from three examining authorities in the years 1994 and 1995. These two years were the last in which the old apartheid style educational divisions were used. The three examining bodies selected were the Department of Education and Training (DET), the Department of Education and Culture of the House of Representatives (HoR), and the Cape Education Department of the House of Assembly (HoA). These three give a fairly representative view of the state of ESL-HG in public education at that

time: the DET catered for African learners, the HoR for 'coloured' learners, and the HoA system for white learners.

In addition, Hansen's research included the first year of the new unitary system, 1996, in which each province examined all the candidates in public schools within its borders. As well as the changes to examining authorities, 1996 saw the implementation of the Interim Core Syllabus, which, as argued above, differed in some important respects from its predecessor, the Core Syllabus, mainly by extending its scope to include ideational and heuristic language uses.

In essence, Hansen concluded that although the overarching aim of the core syllabus on which the ESL examinations were based from 1986 to 1995 was that of communicative competence, the examination papers in these years did not reflect this aim (in other words, the examinations were not adequate in terms of construct representation or relevance). It has been argued at several points in this study that when there is dissonance between the way a subject is assessed and the aims of the syllabus on which instruction is based, attention will be paid to the assessment. As a consequence, the aims of the syllabus will be distorted as teachers attempt to align their teaching with the assessment procedures.

In this section, a brief analysis is undertaken of the ways in which the ESL-HG is examined, in order to explore possible reasons for the inverse relationship of the ELS-HG and the PTEEP tests, as well as the failure of the ESL-HG examination to predict future academic performance in a context where it is the medium of instruction. It needs to be emphasised that the limited and incomplete nature of the documentary data (an inevitable consequence of poor archiving and record-keeping in the DET system in particular, as well as in some of the new examining authorities post-1995) precludes a rigorous or thorough analysis. However, the lack of positive predictive power of the ELS-HG examinations which is consistently indicated in the Yeld et al and Polakow studies[81], coupled with the absence of any significant positive correlation between results on the PTEEP and the ESL-HG examination over any of the years examined, points to a need for some kind of explanatory analysis to be essayed.

[81] These studies are reported in Chapter Ten, sections 10.3.1 and 10.3.2.

Hansen's research led to the following conclusions. For all systems, there was a "... heavy reliance on discrete-point testing of decontextualised grammar items" (Hansen 1997:43). This focus is surprising in light of the fact that the Core Syllabus states explicitly: "A formal programme of work on language structures and usage is neither required nor appropriate" (CS 1986:8), and "However language is used, it should be seen in relation to context: i.e. to purpose, audience, and circumstance" (op cit: 2). Indeed, the Interim Syllabus goes further and suggests that encouraging students to apply what they are learning about language is one of the most effective ways of learning how language works. However, as Hansen (1997:45) suggests, although the communicative aims suggest a central concern for what is successfully communicated, the focus on multiple choice questions militates against this concern.

In relation to textual competence (cohesion and rhetorical organisation), tasks are included which require sentences to be joined using cohesive devices – however, this task occurs only at the level of the sentence (two sentences being combined to form one) and thus provides only a limited picture of candidate ability in this regard. The weighting of textual competence in the writing and literature papers is not possible to assess in the absence of the marking criteria. However, Hansen (op cit:52) argues that "...these papers test mainly content knowledge". Support for this assertion comes from the proliferation of content-based study guides which profess to prepare candidates for the examinations, and which many students read in lieu of the source texts themselves.

One of the most serious of Hansen's conclusions is that "[P]upils ... did not have to understand the texts as a whole or to respond to contextual questions" (op cit:46). The kinds of comprehension tested tended to require either 'content' or 'on the line' comprehension skills, such as those noted by Kapp (2000b).

Understanding of extended text, and literature, was examined entirely though the use of MCQs and some short answer questions. No use was made of extended writing in this regard.

In the main, students in the DET system are expected to either quote directly from passages given in the examination papers, or to complete MCQs. Open-ended questions (or even questions which

go beyond the passages given in the examination papers) are rare. Also, candidates are not required to provide evidence for assertions, to attempt explanations for actions or events, or to comment in any way on 'author-craft' issues. These limitations are serious.

Most of the essay topics are narrative or descriptive. Even those that could be argued to be discursive provide so little material of any substance in the prompt that it is highly likely they would be interpreted by candidates as narrative or descriptive. For example, an essay titled simply "Transport" is presumably intended to elicit a discursive essay on some aspect of transport (e.g. the inadequacy of public transport), but could as easily be interpreted as descriptive.

In relation to the way that the ESL-HG subject is examined, then, it seems that Hansen's primary conclusion is justifiable: the examinations do not reflect the syllabus[82]. While this conclusion assists in explaining the absence of correlation found in the 1988 cohort, it does not assist greatly in understanding why, for these groups of candidates, the two tests exhibit an inverse relationship in both the 1995 and 1996 cohorts. Further investigation is needed to explain this phenomenon. Such an investigation might examine the relationship between the ESL-HG results and other Senior Certificate subjects, the reliability of the ESL-HG results, and the role played of the truncated samples in the Shall and Polakow studies. It is recommended that future PTEEP development cycles incorporate strategies to ensure the gathering of the data that will make such investigations possible.

11.3.3 Similarities and Differences in Predictive Validity

A second, albeit indirect, method of establishing concurrent validity through comparing scores on two tests is to investigate the extent to which the two tests differ in their predictive ability. That is, it is possible to investigate the predictive validity of the one test against (for example) university performance, and to compare this validity to the predictive validity of the other test, against the same criterion performance. The results of this investigation are discussed in some detail in

[82] Hansen (1997:54) suggests the following should be included in tests of communicative competence: "... increasing the weighting of oral and aural competence within the context of authentic communicative tasks; using more open-ended tasks associated with integrative approaches; decreasing the weighting allocated to the testing of discrete grammar

Chapter Ten (Tables 10.2 – 10.6 illustrate these results), and indicate that the two tests differ profoundly in this regard.

The discussion above demonstrates that using the ESL-HG examination as a means of establishing concurrent validity in respect of the PTEEP tests is a problematic undertaking. The discussion revealed that (i) in 1988, no significant relationship was found between the two tests, and that in 1995 and 1996 a significant negative relationship was found, and (ii) that the two tests differ markedly in terms of predictive validity. Essentially, this means that the PTEEP tests cannot be considered to have successfully established concurrent validity on the basis of comparisons with the ESL-HG examinations.

Nevertheless, substantial grounds existed for undertaking a concurrent validation study in respect of these two tests, as both are used for admissions and selection purposes, both are based on broadly similar constructs, and because no other even vaguely comparable test had been written by the same group of candidates as that for whom PTEEP data were available. Indeed, the findings provide important justification for the continued development and implementation of the PTEEP testing initiative, as it is clear that the tests do not simply yield the same information about candidates as do the ESL-HG; and, more importantly, they provide information that, for the group of students on whose performance the study was based, the PTEEP tests are more useful for the purposes of selection and admission.

11.4 Consequential Validity

In the South African context, where the primary goal of the new government is to build a democratic, non-racist, non-sexist society, the consequential validity of tests is a crucial area for concern. Generally speaking, tests and assessment generally are viewed with suspicion. This is not surprising, as their use in the past was frequently linked to some form of discriminatory practice (see, for example, Barsby et al 1994).

---

items; including more tasks, both oral and written, in which discourse and sociolinguistic competence would be tested; and including more tasks which would require students to write discursively".

Messick (1994:19) points out that "[T]here are very few one-edged swords in the measurement enterprise ...", and the accuracy of this observation can be seen in public reactions to some of the unintended consequences of new assessment approaches and techniques. For example, the use of portfolios to elicit and encourage more effective forms of language learning and writing has been challenged on the grounds that they further disadvantage certain groups (e.g. Klein et al 1997). Similarly, it is highly likely that students attending educationally disadvantaged schools which rely on oral learning and do not provide many opportunities for writing (Kapp 2000a&b), will be further disadvantaged, at least in the short term, by being required to take written tests in the performance assessment mode. The SAT in the United States, promoted by its developers as providing a curriculum-neutral alternative to school-based tests that will protect candidates from the disadvantaging effects of poor schooling, has been shown to confirm these effects and further disadvantage learners when their SAT results are used in 'race-blind' admissions procedures[83].

Messick (1989) recommends the use of counterproposals as a useful starting point for the systematic analysis of consequential validity. His specific recommendation in this regard is for test users and developers to assess the potential social consequences of the new or proposed test and its use in relation to the potential social consequences of not testing at all – or of continuing to use existing tests in the case of the development of new test forms or procedures.

In applying this strategy to the development of the PTEEP, the questions need to be slightly reconceptualised, as the tests are already in existence, and so some of the consequences, intended and unintended, are known. The first question is: what are the possible social consequences of the PTEEP tests? The second question is: what would be the social consequences if the PTEEP tests were to be disestablished? The third question, clearly, involves a judgement about which of these sets of consequences is more, or less, desirable.

Geisinger (1992) suggests several important consequences of tests that need to be monitored. These include such matters as possible adverse impact on under-represented groups (see section 2.7.1, Chapter Two, for a discussion of adverse impact and related issues). This criterion is not

[83] This use is not advocated or supported by the SAT developers.

relevant in the PTEEP context, as the project targeted a particular under-represented group. A second consequence could be changes in school or other curricula as educational institutions adapt instruction to a given test. For a non-curriculum-aligned test such as the PTEEP, it is unlikely that teaching to the test would be harmful, as the test construct is based on those general skills believed to underpin academic literacy activities. Nevertheless, this assumption needs to be monitored. Labelling of students (e.g. 'limited English proficient', 'educable retarded', 'slow-stream', or even 'gifted') is another potential consequence. Negative labelling is clearly a danger for any initiative designed to widen access: however, the demand from other groups (i.e. non ex-DET groups) to be included in the scheme suggests that the opportunity to enter UCT outweighed much of the stigma that might have resulted. Closely related to this last is the possibility of disproportionate employment opportunities or assignment to particular - high demand or remedial - curricula. In the PTEEP testing context, with the exception of a single course, no curricula are formally linked to results on the PTEEP tests, and in that case, the results are used to exempt students from, or place them onto, an EAP course (which is a fully credited course within the mainstream menu of programmes in the Humanities Faculty). In terms of disproportionate assignment to a 'remedial' course, this outcome is unlikely to be regarded as an issue – indeed, student evaluations have consistently rated the course as excellent.

Other consequences could include self-assessments by potential candidates, which could result in their excluding themselves from possible opportunities. An example of this (from an admissions testing context) could be that a student, having seen an example of an admissions test from a particular university, decided not to take it as s/he believed she would not 'pass'. In this case the student might decide not to apply for that institution but to go elsewhere. In reality, however, the institution might require quite low scores, which the candidate would in all probability have obtained. In this case, the perceived level of difficulty of the test had the (probably unforeseen) consequence of deterring the student.

Some of the positive consequences of the development and introduction of the PTEEP tests are as follows:

- By the 1999 entry cycle, 1908 students had been admitted to the university at least partly on the basis of their results on these tests[84]. 660 of these students had, by the end of 1999, gained an undergraduate degree, and 104 a postgraduate degree.
- 895 of the 1908 students registered in the fields of science and technology, which are areas where black South Africans are grossly under-represented.
- The numbers of black students at the institution have grown, from 339 at the Project's inception in 1985 to 4,582 in 2000. The growth is not, of course, attributable only to the existence of the tests, but the constant advocacy of the Project, supported by the relatively good academic performance of students admitted via its tests, has undoubtedly contributed to the widening of access opportunities.
- Students' needs have been identified on the basis of their performance in the tests. This outcome has assisted appropriate course and curriculum design, and facilitated placement onto programmes, thereby contributing to a more effective academic experience and environment.

Consequences that could be interpreted as negative include the following:

- The provision of a mechanism whereby UCT is in a position to 'cream off' many of the best black school-leavers from the system, while positive for UCT, could potentially skew access patterns across the system. This needs careful monitoring.
- There is increasing pressure to use the test results as a means of refusing admission. This consequence is positive in that it prevents the demoralising and expensive effects of failure, but it does mean that in effect a further barrier to access has been erected, and in this sense it could be viewed as negative.

Some of the unintended consequences of the tests are identified as follows:

[84] It is difficult to estimate exactly how many students have been admitted as a result of the tests. The numbers of students who have not met cut-off points requirements for various curricula but have been recommended by the project, can be calculated. Approximately one quarter of the AARP recommended students are in this category. It is less straightforward to assess how many candidates who were placed on waiting lists but finally admitted can be attributed to the project. In many cases the fact that they had been recommended by the testing service was the deciding factor, and in others the fact that they had been made provisional early offers on the basis of their PTEEP results (before the SC results were known) meant that they had persisted as candidates and not switched to another institution.

- The tests are used by other institutions to allocate scarce financial resources, or for admissions, or by donors as selection for scholarships and/or bursaries.
- The use of the PTEEP tests in the institution's early offer system, and for financial aid allocation, has meant that educationally disadvantaged students are now considered for early offers, with all the benefits an early offer carries for good planning. Previously, such students were not made early offers, as the schools they attended were unable to provide reliable pre-SC information.
- Increasing use of the test results as a 'canary' for the reliability of the external school-leaving examination system, as well as of school-based examination results.

In evaluating these consequences – that is, in considering whether the positive consequences outweigh the negative ones, it can be argued that in the absence, for the majority of educationally disadvantaged applicants, of useful school-leaving results which institutions can use to admit students, the development and use of the PTEEP and other Project tests has been beneficial in terms of the needs they were designed to meet. That is to say, many hundreds of such students have been admitted to the institution on the basis of their performance on the tests, and a satisfactory number of these students have graduated. Furthermore, to the extent that they provide reliable and useful information about future academic performance at the institution, they can be argued to have had positive consequences even for those students denied admission to the university at least partially on the strength of performance on the tests[85].

Techniques for assessing consequential validity include both statistical and non-statistical approaches. Essentially, they aim to investigate first, whether groups of students (categorised according to sex, ethnicity, age, geographical origin and so forth) perform differently on tests, as a result of some factor in the test rather than as a result of some difference in abilities, and second, what the consequences of these differences are for performances. Chapter Three contains a discussion of test fairness and related issues, and the arguments and caveats are not rehearsed

[85] These positive consequences, are, it must be admitted, long-term, and are unlikely to be acknowledged by the affected individuals. Nevertheless, avoiding the costly (to the student, to the institution, and to the system as a wasted resource) and demoralising effects of failure must surely be considered to be a positive action: the heart of the matter is, however, the degree of confidence in the test results as predictors of performance.

here. From the perspective of consequential validity, however, it is important to note, following Messick (1989), that adverse impact can be attributed to two main sources. The first source arises from test invalidity: that is, some property of the test, such as construct-relevant or -irrelevant variance, or criterion-related or –unrelated variance. In this case the 'problem' is located in the test, and the appropriate response is to redesign or abandon the test. The second source arises from real and relevant differences in the developed abilities of the groups themselves. In this case, the 'problem' is a social and/or educational one, and needs a response such as the provision of appropriate educational opportunities and resources. The fundamental challenge for consequential validity is to ascertain, once differences in performances between groups of students are detected, which source of problem is implicated, so that the response can be appropriately targeted.

The so-called Cleary (1968) model was developed to estimate test bias in relation to prediction, and is presently the dominant statistical approach to test fairness in this regard. Essentially, it compares the regression lines of two groups (e.g. educationally disadvantaged and advantaged students, or men and women) in order to see whether the test as a whole under-predicts or over-predicts their future performance in some specified area, such as first-year university performance. If evidence of bias is found – for instance, if a test under-predicts the future performance of women - a decision can be made to use a different cut-off score for women. For example, they could be admitted with lower scores. Alternatively, the test could be revised until no differences in prediction between groups is found – that is, until the regression lines for both groups do not differ significantly.

As can be seen, the Cleary approach focuses on the whole test as the unit of analysis, and on its relation to some defined criterion situation. Differential item functioning (DIF) methods, in contrast, were developed to investigate whether different groups find individual items within a test more or less difficult than other groups find them: that is, whether different groups fail individual items in disproportionate numbers. In this approach, examinees are grouped according to overall test

score, in order to obtain clusters of ability levels, and then the performance of groups (constituted on the grounds of ethnicity, sex, religion, class) are compared within these ability groupings. DIF methods do not involve performance on some criterion situation, but are focused on intra-test functioning.

Such attempts to establish the existence or otherwise of test bias (that is, to ascertain whether different groups perform differently on a test) were not possible to conduct on the data used in the PTEEP studies reported above. For example, the Cleary approach depends on the existence of two distinct groups writing the same test, and then undergoing some common assessable experience (e.g. studying at university, or taking another test). The aim of the PTEEP predictive validity study, however, was to investigate the degree of confidence that admissions officers could have in the PTEEP results (on their own and in comparison with ESL-HG results) as predictors of academic performance for ex-DET students at UCT. The students are not separated into two distinct groups, and consequently the conditions necessary for legitimate use of the Cleary model did not pertain.

Non-statistical methods for investigating consequential validity could include follow-up studies in schools to see what impact the new test has on teaching and learning preceding it (washback). If a new test claims that it will promote and support the use of metacognitive strategies in problem-solving situations, it would be interesting to investigate, through classroom observation or some form of data collection instrument, whether metacognitive strategy use had indeed increased. It is also important to check for negative consequences. For example, in a context of prolonged reliance on multiple-choice format language tests (e.g. as has been the case with the TOEFL), it is important that the consequences of this practice for the teaching preceding the test be monitored and reported on. The difficulties of relating this recommendation to the PTEEP context are obvious: there are no test preparation manuals, as the test design is based on dynamic principles and sets out to create novel contexts. In addition, the context for the teaching preceding the test is not clearly defined. The point that can usefully be derived from concerns about washback is that

the tests must strive to embody positive models of teaching and learning, and must use a variety of formats.

In terms of consequential validity, then, it can be argued that the introduction of the PTEEP tests has had positive consequences for the target group, that is, for educationally disadvantaged students. It has provided a means of identifying talented students whose Senior Certificate results would not have gained them admission, and thus far has not been the cause of an applicant being refused admission. It has helped the University of Cape Town to meet its goal of rapidly increasing its numbers of black students, while minimising the risk of student failure. Its potential negative consequences, such as reinforcing and exacerbating the 'pecking order', will need to be monitored, as will its possible future use as a means of rejecting very high-risk candidates.

11.4 Conclusion

Chapter Eleven concludes the analysis of external aspects of validity of the PTEEP tests.

In terms of concurrent validity, the analysis reveals that no satisfactory measure of convergent concurrent validity could be established by comparing the PTEEP tests to the ESL-HG examinations (the only available source of appropriate data). However, the analysis demonstrated that the PTEEP was providing different and useful information in its own right.

In terms of consequential validity, the analysis revealed that the introduction of PTEEP tests could convincingly be argued to have had positive consequences.

The following chapter, Chapter Twelve, contains an analysis of the reliability of the PTEEP tests. It concludes the validation study of these tests.

CHAPTER TWELVE

THE PTEEP TESTS AND RELIABILITY

12.1 Introduction

Reliability can be understood as "... the extent to which a test produces consistent results when administered under similar conditions" (Hatch & Faraday 1982:244). This notion is obviously important, not only on grounds of fairness, but also of usefulness. Test users and candidates alike need to have some measure of faith that the score a candidate receives for a test is not random, but is similar to the score s/he would obtain had they written on another day.

Reliability, of course, does not ensure validity: a valid test, however, must be reliable. Reliability is thus a 'necessary but not sufficient' test property - as Nitko (2001:76) suggests, it is "... a limiting factor for validity".

Assessing the reliability of a test can be undertaken in a number of ways, which involve estimating reliability coefficients. Methods of undertaking such estimations fall into two main categories: those which estimate reliability over time, and thus assume some kind of test-retest design; and those which estimate reliability on a single occasion, and are thus based on scores derived from a single administration of a test. Chapter Twelve provides an analysis of approaches to establishing reliability in respect of the PTEEP tests.

12.2 Reliability over Time

Techniques in this category, involve the same test being administered twice, and the calculation of a 'test-retest' reliability coefficient. Several problems are associated with this technique. If the same test is used, and the interval between testing is small, it is difficult to assert that the second performance is not affected by practice or memory (Baker 1997, Zaaiman 1998, Nitko 2001). If the interval is not small, it is difficult to assert that no learning or change in learner state has taken place between the two testing sessions. A variation of this procedure is to employ an alternate form of the first test to use in the second testing session. However, it is not always possible to obtain parallel forms of tests. Some assessment specialists (e.g. Linn & Werts 1979: 55, cited in Baker 1997:7) argue that many so-called parallel test forms are based on "... a series of rather strong assumptions". Indeed, Alderson et al (1995:88) remark that it is "... almost impossible to construct two genuinely parallel tests". Both identical and alternate forms of the test-retest technique suffer, too, from the difficulty of persuading candidates to take the same test with equal seriousness.

Nitko (2001) suggests several reasons why parallel forms of tests are seldom constructed. First, in most cases an assessment procedure will only be used once for each student, so the extra effort is not justified. In the case of the PTEEP, it is virtually impossible to create the space within the hectic first few weeks of a university first-year student's life to undertake any test-retest procedures. Nor is it deemed feasible or legitimate to try to persuade students to retake the test, or a previous PTEEP, during the academic year. Second, the act of taking an assessment may change the test taker in some way, as suggested above, and so parallel forms are not deemed desirable. It is difficult to estimate the extent to which this difficulty applies to the PTEEP test, although the novelty of many of the item formats and topics suggest that candidates might perform differently on a parallel form as they would then be familiar with the approach. In addition, the inclusion of scaffolding makes this issue more complex, as it is premised on the unfamiliarity of many of the test demands – and sets out to 'teach' these. Third, considerations of cost and capacity make it not feasible to develop parallel forms. This is a serious constraint, and highly relevant to the PTEEP context.

The PTEEP testing initiative has thus not adopted a test-retest design to establish reliability.

12.3 Estimating Reliability on a Single Occasion

The difficulties sketched above are not uncommon in test development undertakings. Because of this, many reliability studies make use of techniques which use the scores from one test to estimate reliability. The most usual method of doing this is the split-halves procedure. In this approach, subjects write one test, and then the scores for the test are divided to form two sub-tests. Each subject is assigned two scores, one for each half of the test, i.e. for each sub-test. The two scores are then used as though the subject had taken alternate forms of the same test.

The main difficulty in this method is in splitting the test into two equivalent halves. In the case of a test composed of multiple-choice, discrete items of equal weight, division into odd and even numbers is a workable approach. However, difficulties arise when - as is the case with the PTEEP tests - items are clustered into homogeneous groups, or stem from the same stimulus material (e.g. from a table, graph, or particular text). These difficulties are discussed below in relation to the PTEEP tests.

12.4 Assessing the Reliability of the PTEEP Tests

In assessing the reliability of the PTEEP tests, the 'split-halves' method was adopted for parts of the tests, as follows. First, only items that could be objectively marked were included (that is, true/false, multiple-choice, and one-word answer items). The items were allocated to 'test 1' or 'test 2', as far as possible in strict order of occurrence. However, this approach is not appropriate when items are linked on some basis, or when the design of the test deliberately sets out to build skills, thereby making some of the answers contingent in complex ways on each other. In assigning items to the two sub-tests, judgement was exercised and inappropriate items were not assigned to either test. Nevertheless, it cannot be asserted with confidence that the two subtests were directly comparable, and therefore the application of the split-halves approach is of questionable value.

The results of the split-halves method applied to the 1998 and 1999 PTEEP tests (item-level data had not been entered for the earlier tests) are displayed in Table 12.1 below.

	1998 PTEEP (N = 1,018)	1999 PTEEP (N = 989)
No. of items included in 'split- halves' procedure	50	30
Split-halves reliability index	0.85	0.68
Spearman Brown double-length formula	0.92	0.81
Cronbach's alpha (coefficient alpha)	0.77	0.86

Table 12.1: Reliability of the 1998 and 1999 PTEEP tests

According to the American Office of Educational Assessment (1998), reliability coefficients are dangerous to interpret out of context. As a rule of thumb, however, they suggest that coefficients of 0.9 and above can be regarded as excellent, and of 0.8 and above as highly satisfactory. The two tests for which data were available can thus be argued to display satisfactory properties of reliability, although several important caveats exist.

The Spearman-Brown double-length formula adjusts the correlation that results from the correspondence of two sets of scores that are far smaller than (strictly speaking, half the size of) the whole test, to represent an estimation of the reliability of the whole test. However, the Spearman-Brown formula is not usually regarded as appropriate to use with tests such as the PTEEP, for the following reasons. First, the scores used in the rank order and Spearman-Brown double-length procedures refer to only part of each test, and not to the whole, which includes essays and other constructed response items. The greater number and proportion of items suitable for inclusion from the 1998 test compared to the 1999 test might help to explain why the coefficient for the 1998 intake test is higher than that for the 1999 test. Second, the items are not independent of each other, as some items build on answers to previous items. Finally, as mentioned above, the design of the test rests on the notion of task scaffolding, and so it is anticipated that certain items will be contingent on others – and some will be easier or more difficult. For this reason, the analysis needs to be viewed with considerable caution.

Similar reservations exist about the suitability, in this context, of the Kuder-Richardson formulas 20 and 21, which are appropriate only for dichotomously scored items only. For this reason, the study

calculated, in addition, coefficient alpha (Cronbach's alpha), which can be used with items that are dichotomously scored (i.e. the student gets one of two scores – usually 1 or 0) or polytomously scored (i.e. where a range of scores is possible), as well as with tests containing heterogeneous tasks (i.e. tasks measuring more than one trait).

It can be seen from Table 12.1 that the parts of the tests included in the reliability analysis exhibit acceptable reliability coefficients. The 1999 PTEEP, particularly, with a coefficient alpha of 0.86, can be regarded as satisfactory in this respect.

It is important that traditional approaches to establishing reliability, such as those reported above, are followed, and the results reported in the PTEEP test manuals. However, the nature of the tests described and analysed in this study suggests that results from these traditional approaches will need to be interpreted with great caution. For this reason, it is imperative that the PTEEP development process takes whatever steps are possible to improve reliability in the test design itself. Hughes (1989) suggests that these steps include the following:

- Taking enough samples of behaviour – that is, maximising the number of items;
- Restricting the choices within a test, thereby enhancing comparability and uniformity;
- Writing unambiguous items, with clear and explicit instructions;
- Paying attention to the lay-out of tests;
- Exercising due diligence in respect of new formats. This could be achieved by careful piloting and by ensuring that the skills and abilities that are intended to be elicited by items utilising new format should also be elicited by tried and trusted formats.
- Standardising administrative procedures as far as possible;
- Maximising objectivity in terms of scoring, for example by using sentence completion tasks rather than, or in addition to, tasks which require whole sentences to be written;
- Providing detailed scoring keys and training scorers thoroughly on their use;
- Ensuring anonymity of scripts so that marker bias is minimised; and
- Employing multiple, independent scoring techniques, especially of subjective items.

As was demonstrated in the foregoing chapters, and in particular in Chapter Nine in which the content validity of the PTEEP tests was investigated, substantial efforts were made to ensure that these suggestions were incorporated in the design and analysis of the tests. The thoroughness of the procdures employed in the design stage suggests that the tests are likely to be reliable: however, as has been discussed in the study at various points, the data entry procedures in particular need to be professionalised in order to facilitate as far as possible the investigation of the tests' reliability. It is recommended that this be done for future development cycles.

12.5 Conclusion

It has been shown above that the PTEEP tests exhibit an acceptable level of reliability insofar as it is possible for this to be demonstrated given the fundamental incompatibility of the testing approach with classical test theory. Nevertheless, it was argued that it was necessary for traditional approaches to be followed even if serious caveats constrained the interpretation of the results. In addition, the chapter highlighted the importance of various design features and procedures in enhancing reliability, and it was argued that, on the evidence of the preceding chapters, this importance had been recognised.

Chapter Twelve concludes the investigation into the validity and reliability of the PTEEP tests. The investigation aimed to address the second major research question identified in Chapter One, section 1.2, viz. To *what extent are the PTEEP tests, developed to identify talented but educationally disadvantaged candidates whose SC results would not necessarily reveal their abilities, valid in terms of the construct established at the end of Chapter Six?*

To recapitulate briefly, Chapter Six concluded by putting forward a construct for a contextually appropriate test of academic literacy (that is, appropriate in a context of widespread educational disadvantage where considerations of cost are important) that could be used as an additional source of information for the purpose of selection for Higher Education. Chapters Seven to Twelve then set out to assess to what extent the PTEEP tests, developed to broaden effective access

opportunities for educationally disadvantaged, black students to the University of Cape Town, could be considered to be valid in terms of the construct laid out in Chapter Six.

In short, the PTEEP tests were argued to be valid and reliable in many respects. These are detailed in Chapters Eight to Twelve. Claims to validity are supported with evidence, where this was appropriate and where data were available, and/or through discussion of test development procedures in other cases. As was made clear particularly in Chapter Seven, however, the PTEEP tests were developed in the early days 'on the hoof', and until relatively recently data entry and archiving procedures were not conducive to rigorous validation scrutiny. The study has made clear recommendations on how procedures in these regards can be improved.

In relation to the research question, it can therefore be argued that the PTEEP tests are, to a considerable extent, valid in terms of the construct outlined at the end of Chapter Six, although a more definitive assessment is not possible owing to some important lacunae in terms of data. Nevertheless, in respect of critical aspects of the construct (such as the incorporation of dynamic testing principles, and the predictive aim of the tests) the PTEEP tests were shown to be valid.

Chapter Thirteen draws the study to a close by summarising the findings, making several recommendations on the basis of these, and putting forward some suggestions regarding implementation.

CHAPTER THIRTEEN
CONCLUSIONS AND RECOMMENDATIONS

13.1 Conclusions

The central problem identified at the outset of the study was that, for the majority of applicants to Higher Education, the current Senior Certificate results are not a useful predictor of future academic performance. For this same majority, the language of learning (the language in which the Senior Certificate examination is written and through which they will study at tertiary level) is not their first language. In addition, and for this same group of applicants, the quality of schooling is highly likely to have been of poor quality. These two factors (language and poor schooling) interact in complex ways and contribute powerfully to what has become known as educational disadvantage.

Nevertheless, Higher Education institutions in South Africa need to be able to select from amongst this large majority, and the Senior Certificate is the only academic achievement indicator that is comparable across candidates.

It is in this context that the aims identified in Chapter One arose. In addressing these, a wide range of fields of study was covered, and a number of methodological approaches employed. In summary, the aims of the study were first, to identify methods of selection that could be used in addition to Senior Certificate results, and to identify ways in which effective access for educationally disadvantaged applicants can be widened. Second, the study set out to assess the extent to which the academic literacy tests developed by the Alternative Admissions Research Project at the University of Cape Town could be considered to be valid in terms of these methods and ways. In addition, while this was not the primary focus of the study, it aimed to assess the

extent to which the English Second Language Higher Grade (ESL-HG) examination could be considered to be promoting the development of cognitive academic language proficiency.

The importance of the Senior Certificate examination in the South African schooling system was repeatedly emphasised in the study. This importance relates both to its role as quality assurer in the virtual absence of other mechanisms at this level, and to its motivational role for learners and educators alike, although the negative aspects of high stakes assessment regimes were noted. The identification of other methods of selection was thus proposed to provide a source of additional information rather than to challenge the Senior Certificate system.

This proviso means that any additional information cannot be derived from an achievement test based on school subject areas, as this would lead to inevitable comparisons and unhelpful competition. Besides the likelihood of a similar type of examination yielding similar results to those already available, considerations of cost and capacity make the duplication of effort involved in mounting a parallel examination not feasible. As was argued in the dissertation, any such efforts should be directed at strengthening and improving the Senior Certificate itself rather than creating a competitor.

In Chapter Two, alternatives to selection (for Higher Education) were discussed, and it was concluded that selection was unavoidable in contexts of excess demand, as well as in the best interests of society. The discussion then turned to the vexed question of the development of fair criteria on which to base selection decisions. It was argued that whatever model for selection is eventually adopted, the need to predict academic performance will remain. Determining academic merit was recognised, however, to be a particularly complex and controversial matter in a society with as high a degree of inequality of educational provision as that of South Africa. In this context, it was concluded that reliance on curriculum-aligned achievement tests such as the Senior Certificate would discriminate against candidates whose schooling had not been of high quality. The chapter thus concluded that additional measures needed to be investigated.

Chapter Three embarked on this investigation by providing a comprehensive overview of major assessment issues. The strengths and limitations of the 'assessment-led-instruction' school of thought were discussed, as were the criteria governing appropriate test use. Several issues were highlighted which would need to be addressed in any proposal emanating from this study. First, tests that bear a close relationship to a course of instruction (curriculum-aligned tests) would unfairly discriminate against those whose opportunities to learn have been poor. However, tests that do not have a close relationship to a course of instruction would, it was argued, inevitably distort the curriculum if used in isolation.

The chapter therefore concluded that since reliance on either one of these types of test would have negative consequences in a context where educational inequities are known to exist, both types are needed, to be used in combination.

Chapter Four argued the case for a principled, progressive use of assessment in selection procedures to Higher Education. The argument was made that in contexts of extreme heterogeneity of educational provision, reliance on static methods of assessment for selection purposes, whether the tests are curriculum-aligned or non-aligned (the focus of Chapter Three), is neither fair nor useful. Such reliance would not be fair, because educationally disadvantaged students will inevitably perform very poorly in competition with their more advantaged peers. In addition, reliance on static methods alone would not be useful for selection purposes, in that educationally disadvantaged students produce a restricted range of scores at a very low level.

The chapter therefore reviewed various approaches to what has become known as dynamic assessment, which attempt, however approximately, to explore what a candidate could have achieved had s/he had appropriate learning opportunities, rather than simply what s/he has achieved. In Vygotskian terms, this can be described as attempting to tap the Zone of Proximal Development and not only the Zone of Actual Development (Vygotsky 1978).

The chapter concluded that non curriculum-aligned, core skills tests that are developed as far as possible on dynamic lines represent the most effective and fair approach to assessment in the

South African context, when used in combination with the Senior Certificate examination which yields important information about what candidates have actually learned. It was further concluded, on the basis of the review, that scaffolding holds promise as a dynamic assessment technique.

Thus far, then, the study had demonstrated that (i) the Senior Certificate ought not to be the sole indicator of academic merit, (ii) non-curriculum aligned tests appear on the surface to provide an additional source of information, (iii) educationally disadvantaged students perform very poorly on 'static' tests, even when they are non-curriculum aligned, and (iv) dynamic testing could add the missing dimension.

The decision to avoid subject or discipline areas, reported above, has important consequences for test design. One of the main advantages – for test developers - of curriculum-aligned achievement tests is that the domain to be assessed is determined by the content (knowledge and skills) of the course of instruction preceding it. Tests that are not based on a curriculum, however, have to establish a basis, or domain. In this study, the criterion situation was that of Higher Education, and thus the domain of necessity needed to comprise the core skills and abilities employed in this context. Following trends and established practice elsewhere, the assessment of core skills in the area of academic literacy was investigated. Various difficulties and inherent tensions in this area were discussed, in particular the importance of discipline or domain expertise in the development and demonstration of core skills and abilities.

Chapter Five explored and analysed major approaches to knowing and learning in the context of academic literacy, and emphasised the challenges for large-scale assessment that are inherent in the testing of higher-order cognitive skills[86] in particular. The chapter concluded that the most effective approach to gaining insight into candidates' levels of functioning in this regard is to include as wide a range of item formats, and as great a number of items, as possible in the testing

[86] Insights would need to be gained on such matters as the extent to which learners' understanding in a domain demonstrates integration (for example, between items of information, skills, situations, and types of problem), and a generative capacity, enabling knowledge and skills learned or encountered in one situation or for one purpose to be applied in or to another.

of any one particular skill or ability. It was argued that this would make it possible for different representations of the skill or ability to be explored[87].

Chapter Six focused on the role of language and language testing in developing and assessing academic literacy. Insights and conclusions from earlier chapters were drawn on in order to establish an appropriate construct for an academic literacy test. Notions of how learning takes place, how it can be promoted or retarded, and what it is to know and use a language in academic contexts, were incorporated into a set of principles underlying a test specification blueprint.

Chapters Two to Six laid the conceptual foundations for the study, and concluded by setting up blueprint, or construct, for an appropriate academic literacy test in this context.

The second part of the study took the form of a validation assessment of the PTEEP tests which had been developed to meet the needs identified at the beginning of this study (i.e. to identify talented but educationally disadvantaged candidates whose Senior Certificate results would not necessarily reveal their ability). The validation exercise assessed the extent to which the tests could be considered to be valid in terms of the construct developed in Chapters Two to Six, and articulated in Chapter Six. The aspects or kinds of validity that were investigated were construct, content, face, response, predictive, and concurrent.

Chapter Seven traced the origins and history of the development of the Alternative Admissions Research Project at UCT, in order to provide a context for the validation study of the PTEEP tests.

The construct validity of the PTEEP tests was assessed in Chapter Eight. It was concluded that the tests are valid in many respects, in terms of this aspect of validity. The PTEEP construct maintains that it is through the incorporation of dynamic testing approaches that the abilities of educationally disadvantaged students can most effectively be revealed. Section 8.4.2 investigated this claim by analysing student performance on two tests, which, while comparable in most

[87] This was recognised as essential not only for gaining realistic and useful knowledge about the candidates, but also in terms of the 'washback' effect of large-scale testing, in that ways in which such tests are constructed affect the teaching that goes before them, even with non curriculum-aligned tests.

respects, differ in respect of incorporating scaffolding as a dynamic testing approach technique. That is to say, one test incorporated scaffolding, while the other did not. It is demonstrated, through a comparison of performance on items that were identical in both tests, that the new approach produced a greater range of scores; increased predictive validity; and raised the level of the stronger students' scores.

It was concluded that the use of scaffolding within a test, for talented educationally disadvantaged students, can significantly enhance test performance.

It was also concluded that the inclusion of a wide variety of tasks that require candidates to engage with the test content in a number of different ways can be considered to have provided effective ways in which candidates could reveal their abilities.

Chapter Nine was devoted to an analysis of content, face and response validity. Claims to validity in these respects are largely based on design and procedural features, and not on the basis of statistical or quantitative methods. Discussion thus focused on the ways in which the development of the tests could be considered to have met generally agreed requirements in the identified areas.

In terms of content validity, the following conclusions were reached:

- the tests are valid insofar as construct representation is concerned. This was argued largely on the basis of the comprehensive coverage of the specifications arrived at by the cross-disciplinary panel of experts constituting the test development team for each test.
- the tests are valid insofar as construct relevance is concerned. A componential model of task characteristics (comprising situation, material, and rubric, themselves further subdivided) was used to assess the extent of validity in this respect.

In terms of face and response validity, the study concluded that it was not possible to assess the extent to which the tests could be considered valid in these respects, as the PTEEP testing initiative had not undertaken the kinds of studies (e.g. interviews or surveys) which would allow

such assessments to be made. It was recommended that these be undertaken in future development cycles.

Chapter Ten addressed the aspect of validity that is perhaps most critical for an admissions test - that of predictive validity. Two kinds of predictive validity studies were reported; correlational and survival-analysis studies. In respect of correlational studies, it was concluded that although interesting and important information was yielded, the studies failed to provide the kinds of information about any of the hypothesised predictor variables (PTEEP test scores, Senior Certificate point scores and ESL-HG scores) that would be useful for selection purposes. Where predictive relationships were shown to exist, they were weak, and revealed little that was useful from an admissions officer's point of view about how candidates would progress through their degrees.

A different approach to investigating predictive validity, known as 'survival analysis', was therefore adopted. This technique makes it possible to deal with one of the central challenges of cohort studies: how to analyse the data while many of the subjects are still continuing with their studies. The study was based on the observed exclusion rates of various groups of ex-DET students.

Several interesting and important findings emerged from the survival analysis study conducted by Polakow. For example, serious problems were revealed about the use of first-year performance as an indicator of subsequent performance. In addition, it was shown that the group of students who volunteer to write the AARP tests have significantly lower Senior Certificate results than the group that does not write the AARP tests.

Most importantly, it was concluded that, for the group of students in the study, the PTEEP tests are effective in terms of predicting academic success at the University of Cape Town. Specifically, the study demonstrated that students who score in the top quintile of their candidate pool are less likely to be excluded (and thus are more likely to graduate) than are students who are admitted on the basis of their Senior Certificate results alone.

It was further concluded that poor performance on the PTEEP tests should be taken seriously. This was based on the finding that the bottom quintile of PTEEP performers have very significantly higher rates of exclusion than do students admitted on the basis of their Senior Certificate results alone[88]. This suggests that even if applicants obtain Senior Certificate scores that meet regular admission requirements, very poor performance on the PTEEP test reveals a potential problem.

On the basis of these findings, it was concluded that the PTEEP tests can be considered to be valid in terms of predictive validity.

The final aspect of validity that was investigated in respect of the PTEEP tests was that of concurrent validity. The ESL-HG examination was identified as the only possible test for this purpose, as all the students in the study had written both it and a PTEEP test. In addition, its construct is based on very similar notions of academic literacy as the PTEEP tests. However, significant differences in performance on the two tests emerged in the concurrent study. Where significant relationships were shown to exist, these were negative.

The investigation suggested that the ESL-HG examination could not be considered to be valid in terms of its construct, as defined by the syllabus. It was recommended that further study into the examining of ESL-HG be undertaken.

In terms of concurrent validity, it was concluded that the PTEEP tests exhibited divergent validity in that there was no positive relationship between the two. Convergent validity was not established. As was discussed in the study, this is a positive finding for the PTEEP tests, in view of the inverse or random correlation between performance on the ELS-HG examination and performance at UCT reported in Chapter Eight.

Regarding reliability (Chapter Twelve), it was concluded that traditional approaches to investigating test reliability were inappropriate given the use of scaffolding in the PTEEP test design. Although satisfactory levels of reliability were obtained using coefficient alpha and the Spearman Brown

[88] At the time of this study, no student was refused admission on the basis of their PTEEP scores - if they met Senior

double-length formula, it was concluded that it was difficult to interpret their meaning. Because of the incompatibility of traditional reliability approaches with the PTEEP test construct and approach, it was noted that great care needed to be taken at the design stage of the tests to ensure that reliability is maximised, and several steps were discussed in this regard. It was concluded that the PTEEP tests could be regarded to have taken the required steps.

The need for further research is emphasized repeatedly in the study. In addition to the obvious need for meticulous and detailed data entry procedures, the study recommends that:

- efforts be made to investigate the cognitive aspects of test taking in respect of tasks and items. Methodological approaches here include such techniques as think-aloud protocols, and aim to provide insights into the cognitive demands of various item types. Such undertakings would greatly enhance test developers' understandings of the role of scaffolding within tests.
- investigations relating to face and response validity need to be undertaken. Such approaches (e.g. interviews or surveys) would fill an important gap in understanding the role and impact of PTEEP-type tests.
- wherever possible, efforts be made to balance quantitative with qualitative approaches, and in particular, that data entry procedures be comprehensive and meticulous.

13.2 Recommendations

The recommendations put forward here relate to educationally disadvantaged students only. This restriction is a consequence of the limitations of the study, which investigated the validity of the PTEEP tests in respect only of educationally disadvantaged students.

Since educational disadvantage is a social and not a 'racial' construction, it is essential for it to be carefully defined. The definition that is proposed is that put forward in the report of the Committee on the Senior Certificate Examination (DoE 1998). In the report, educationally disadvantaged students are defined in order to make it possible for the "...adjusted marks of their non-language

---

Certificate admissions requirements, they were offered a place.

subjects to be multiplied by 1,05" (op cit:28). This has been done since the 1998 Senior Certificate cycle, and so the definition is clearly workable.

Educationally disadvantaged students are defined in the report as "... candidates who offer an African language[89] as a First Language in the Senior Certificate examination, and not also English or Afrikaans as a First Language ..." (DoE 1998:28, emphasis in original).

The reasons given in the report for this definition are first, that "... the teaching and assessment of African First Languages is aimed at language maintenance, rather than at cognitive development as is the case with English [and Afrikaans] First Language. ...[t]hese differences seriously disadvantage students who study an African language as a first language and who do not study either Afrikaans or English First Language" (op cit:45).

In addition, the committee argues that "[I]t is clear that candidates who write their non-language subjects in a language other than their first language, are at a considerable disadvantage not only in terms of their own performance (i.e. whether they can fully understand the questions and convey their understanding to the examiners effectively) but also because they are being assessed in comparison with other candidates who are writing in their first language" (op cit:28). These disadvantages apply equally to the situation of candidates writing the proposed PTEEP-type tests, which are 'strong' language tests: that is, tests where language is the vehicle by means of which candidates demonstrate their ability on a range of tasks, rather than the target (where language is the end in itself).

Finally, the definition is believed to include the overwhelming majority of all students studying at schools that would previously have been categorised as DET.

It is recognised that some students not included in the definition above could legitimately be considered as disadvantaged, and it is recommended that institutions allow some flexibility in this regard. For example, candidates from some ex-HoR schools, which cover an extremely wide range of preparation and language background (Morris 1985), as well as some candidates who

have transferred only in the last two years of schooling from ex-DET to advantaged schools, could be considered as disadvantaged.

The recommendations are as follows.

13.3.1 It is recommended that Higher Education institutions include, as part of their selection criteria, and in addition to Senior Certificate results, a test which is:

- non curriculum-aligned. That is, the test should not be based on any particular course of instruction.
- based on the domain of academic literacy as defined in the construct in Chapter Six.
- developed on the basis of dynamic principles. The approach adopted in the PTEEP tests validated in this study used the technique of 'scaffolding' as its method of incorporating dynamic principles.

As a convenient shorthand, tests developed along these lines are here called PTEEP-type tests.

13.3.2 It is recommended that excellent performance on these tests be viewed as a legitimate means of selection, equal to that of the Senior Certificate. That is, that despite poor performance on the Senior Certificate, candidates should be accepted on the basis of their PTEEP-type test results. This should not be taken as being in contradiction with the repeated emphasis in this study on the importance of Senior Certificate results. On the contrary, the results of a candidate with poor Senior Certificate results but excellent PTEEP-type test results should be used as a guide to place students appropriately, for example onto foundation courses or extended curricula.

While the Polakow survival analysis study suggests that poor performance on PTEEP-type tests can be taken as an indicator of a high degree of future academic risk, however, it is strongly recommended that further evidence should be amassed, and the consequential implications seriously considered, before students are denied admission on this basis.

[89] The nine official languages, other than English and Afrikaans, were known in the report as African languages.

13.3.3 It is recommended that as a matter of urgency, research be conducted into the examining of ESL-HG. The research should be focused particularly on the extent to which the examination does and/or could promote the development of CALP skills.

13.3.4 It is recommended that the Department of Education give serious consideration to the potential of such a test to strengthen its Quality Assurance efforts at the school-leaving/Higher Education entry interface. If this recommendation were accepted, the test could be written as an extra paper in the Senior Certificate examination set, in order to provide a full picture of the levels of core skills at this level. The recommendations in 13.3.2.1 and .2 above refer only to candidates applying for Higher Education, and would result in a highly skewed sample. Its use for Quality Assurance would thus be compromised unless all SC candidates were involved as is suggested here.

13.3.5 Regarding implementation, it is recommended that:

- the costs of the test should be shared between the DoE, individual institutions, and candidates, in the scenario where individual institutions require the writing of the test. The subsidy from the DoE is motivated on the grounds that improved yet equitable selection procedures would improve the efficiency of the system (through-put rates, effective placement etc.); and would provide valuable information for planning and implementation for the system.
- The tests should be developed and resulted by an independent body commissioned by Higher Education governance structures (e.g. the South African Universities Vice-Chancellor's Association), in consultation with the Department of Education.

REFERENCES

Adelman, C. (1999). Why Can't We Stop Talking About the SAT? *The Chronicle of Higher Education*, November 5, 1999, B4-5.

Adelman, C. (Ed.)(1988). *Performance and Judgment: Essays on Principles and Practices in the Assessment of College Student Learning*. Washington, DC: Office of Educational Research and Improvement.

Ajayi, J.F.A., Goma, L.K.H. and Johnson, G.A. (1996). *The African Experience with Higher Education*. Ghana: The Association of African Universities.

Alderson, J.C. (1988). New Procedures for Validation Proficiency Tests of ESP? Theory and Practice. *Language Testing* 5(2), 220-232.

Alderson, J.C. (1981a). Introduction. In Alderson, J.C. and Hughes, A. (Eds.) (1991). *ELT Documents 111 – Issues in Language Testing*. London: The British Council, 5-8.

Alderson, J.C. (1981b). Reaction to the Morrow Paper. In Alderson, J.C. and Hughes, A. (Eds.)(1991). *ELT Documents 111 – Issues in Language Testing*. London: The British Council, 45-54.

Alderson, J.C. (1980). Native and Non-Native Speaker Performance on Cloze Tests. *Language Learning* 30(1), 59-76.

Alderson, J.C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C. and Clapham, C. (Eds.)(1992). *Examining the ELTS Test: An Account of the First Stage of the ELTS Revision Project*. (IELTS Research Report No. 2). Cambridge, England: The British Council/University of Cambridge Local Examinations Syndicate.

Alderson, J.C. and Hughes, A. (Eds.)(1991). *ELT Documents 111 – Issues in Language Testing*. London: The British Council.

Allard, F. and Burnett, N. (1985). Skill in Sport. *Canadian Journal of Psychology* 39, 294-312.

Altbach, P.G., Arnove, R.F. and Kelly, G.P. (Eds.)(1982). *Comparative Education*. New York: MacMillan.

Altink, W.M.M. and Thijs, G.D. (1984). The Issue of Equity: Research on Selection Processes for Educational Programmes in Developing Countries. *IDS Bulletin* 15(4). Sussex: Institute of Development Studies.

Amano, I. (1990). *Education and Examination in Modern Japan*. Tokyo: University of Tokyo Press.

Amos, T.L. (1999). Integrating the Development of Academic Literacy into Mainstream Teaching and Learning. *South African Journal of Higher Education* 13(3), 177-183.

Anastasi, A. (1982). *Psychological Testing*. New York: Macmillan.

Arena, L.A. (1975). *Linguistics and Composition: A Method to Improve Expository Writing Skills.* Washington, DC.: Georgetown University Press.

Babad, E. and Budoff, M. (1974). Sensitivity and Validity of Learning Potential Measurement in Three Levels of Ability. *Journal of Educational Psychology* 66, 439-447.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L.F. and Cohen, A.D. (Eds.)(1998). *Interfaces Between Second Language Acquisition and Language Testing Research.* New York: Cambridge University Press.

Bachman, L.F. and Cohen, A.D. (1998). Language Testing – SLA Interfaces: An Update. In Bachman, L.F. and Cohen, A.D. (Eds.)(1998). *Interfaces Between Second Language Acquisition and Language Testing Research.* New York: Cambridge University Press, 1-31.

Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice.* Hong Kong: Oxford University Press.

Badsha, N., Blake, G.T.W. and Brock-Utne, J.G. (1986). Evaluation of the African Matriculation as a Predictor of Performance in the University of Natal Medical School. *South African Journal of Science* 82, 220-221.

Badsha, N., Griesel, H., Smith, M. and Yeld, N. (Eds.)(1992). *Proceedings of Cintsa Admissions Symposium.* Cape Town: University of Cape Town.

Badsha, N., Williams, A. and Yeld, N. (1987). *Alternative Admissions Research Project: A Work in Progress Report.* Paper presented at the Academic Support Programme Annual Conference, Grahamstown, December 1987.

Baker, R. (1997). *Classical Test Theory and Item Response Theory in Test Analysis.* Special Report No 2: Language Testing Update. Lancaster, Lancs.: Lancaster University, UK.

Ball, C. and Eggins, H. (Eds.)(1989). *Higher Education into the 1990s: New Dimensions.* Milton Keynes: Society for Research into Higher Education and Open University Press.

Barsby, T., Haeck, W. and Yeld, N. (Eds.) (1994). *Accountability in Testing: the Contribution of Tests to Increased Access to Post Secondary Education.* South African Association for Academic Development - Workshop Publication Series No.2. Cape Town; University of Cape Town.

Barton, D., Hamilton, M. and Ivanic, R. (Eds.)(2000). *Situated Literacies: Reading and Writing in Context.* London: Routledge.

Barton, P. (1994). *Becoming Literate about Literacy.* Princeton, NJ: Educational Testing Service.

Bartram, D. (1995). Predicting Adverse Impact in Selection Testing. *International Journal of Selection and Assessment* 3(1), 52-61.

BBC News/UK Systems (June 10 1999). Oxford Seeks Fairer AdmissionsTests. http://news.bbc.co.uk/hi/english/education/uk_systems/ default.stm (accessed 15 May 2000).

Beatty, A., Greenwood, M.R.C. and Linn, R.L. (Eds.)(1999). *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions*. Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington DC: National Academy Press.

Bereiter, C. and Scardamalia, M. (1982). From Conversation to Composition: the Role of Instruction in a Developmental Process. In Glaser, R. (Ed.)(1982). *Advances in Instructional Psychology, Volume 2*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bialystok, E. (Ed.)(1991). *Language Processing in Bilingual Children*. Cambridge: Cambridge University Press.

Bialystok, E. and Sharwood-Smith, M. (1985). Interlanguage is not a State of Mind: an Evaluation of the Construct for Second Language Acquisition. *Applied Linguistics* 6(2), 101-117.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*. White Plains, NY.: Longman.

Bowen, W.G. and Bok, D. (1998). *The Shape of the River*. Princeton, NJ.: Princeton University Press.

Bonanno, H. and Jones, J. (1997). *Measuring the Academic Skills of University Students*. Sydney: University of Sydney Learning Assistance Centre Publications.

Bot, M. (1994). A Brief Overview of Education, 1993. *EduSource* 5, 1-12.

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. (R. Nice, Trans.). London: Routledge and Kegan Paul.

Bradbury, J., Damerell, C., Jackson, F. and Searle, R. (undated). ESL Issues Arising from the "Teach-Test-Teach" Programme. Unpublished manuscript. Durban: University of Natal.

Bransford, J.D., Declos, V.R., Vye, H.J., Burns, M.S. and Hasselbring, T.S. (1987). 'State of the Art and Future Directions'. In Lidz, C.S. (Ed.)(1987). *Dynamic Assessment: An Interactional Approach to Evaluating Learning Potential*. London: Guilford Press, 479-496.

Breier, M., Taetsane, M. and Sait, L. (1996). Taking Literacy for a Ride – Reading and Writing in the Taxi Industry. In Prinsloo, M. and Breier, M. (Eds.)(1996). *The Social Uses of Literacy: Theory and Practice in Contemporary South Africa*. Johannesburg: Sached and John Benjamin Publishing Company, 213-233.

Brown, A.L., Campion, J.C. and Day, J.D. (1981). Learning to Learn: On Training Students to Learn from Text. *Educational Researcher* 10, 14-21.

Browne-Miller, A. (1995). *Intelligence Policy. Its Impact on College Admissions and Other Social Policies.* New York: Plenum Press.

Budoff, M. and Friedman, M. (1964). "Learning Potential" as an Assessment Approach to the Adolescent Mentally Retarded. *Journal of Consulting Psychology* 28, 433-439.

Bunting, I. (1994). *A Legacy of Inequality: Higher Education in South Africa.* Cape Town: University of Cape Town Press.

Camara, W.J. and Schmidt, A.E. (1999). Group Differences in Standardised Testing and Social Stratification. *College Board Report No. 99-5.* New York: The College Entrance Examination Board.

Canale, M. (1983a). From Communicative Competence to Communicative Language Pedagogy. In Richards, J.C., and Schmidt, R.W. (Eds.)(1983). *Language and Communication.* London: Longman, 2-27.

Canale, M. (1983b). On Some Dimensions of Language Proficiency. In Oller, J.W. (Ed.)(1983). *Issues in Language Testing Research.* Rowley, MA: Newbury House, 333-342.

Canale, M. and Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics* 1(1), 1-47.

Carrell, P.L., Devine, J. and Eskey, D. (Eds.)(1988). *Interactive Approaches to Second Language Reading.* Cambridge; Cambridge University Press.

Carrell, P.L. (1988). Some Causes of Text-Boundedness and Schema Inference in ESL Reading. In Carrell, P.L., Devine, J. and Eskey, D. (Eds.)(1988). *Interactive Approaches to Second Language Reading.* Cambridge; Cambridge University Press, 101-113.

Carroll, J.B. (1961). Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. In Center for Applied Linguistics (1961). *Testing the English Proficiency of Foreign Students.* Conference Proceedings. Washington D.C.: Center for Applied Linguistics.

Chang, M.J. (1999). Does Racial Diversity Matter? The Educational Impact of a Racially Diverse Undergraduate Population. *Journal of College Student Development* 40(4), 377-395.

Chi, M.T.H. (1978). Knowledge Structure and Memory Development. In Siegler, R.S. (Ed.)(1978). *Children's Thinking: What Develops?* Hillsdale, NJ: Erlbaum, 73-96.

Chi, M.T.H., deLeeuw, N., Chiu, M. and LaVancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science* 18, 439-477.

Chi, M.T.H., Feltovich, P.J. and Glaser, R. (1981). Categorisation and Representation of Physics Problems by Experts and Novices. *Cognitive Science* 5, 121-152.

Chilisa, B. (1999). New Developments in the National Examination System in Botswana. *Educational Measurement: Issues and Practice* 18(4), 28.

Chomsky, N.(1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clapham, C. (1996). *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*. Cambridge: Cambridge University Press.

Cleary, T.A. (1968). Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement* 5(2), 115-1254.

Cloete, N. and Pillay, S. (1987). *Can Wits Have a More Representative Student Body Within the Existing Structures: Optimistic or Pessimistic?* Unpublished Manuscript. Johannesburg: University of the Witwatersrand.

Cohen, A.D. (1998). Strategies and Processes in Test Taking and SLA. In Bachman, L.F. and Cohen, A.D. (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*. New York: Cambridge University Press, 90-111.

Cohen, I.S. (Ed.)(1989). *The G. Stanley Hall Lecture Series* 9. Washington, DC.: American Psychological Association.

Cohen, L. and Manion, L. (1994). *Research Methods in Education (4th Edition)*. London: Routledge.

Coleman, P., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. and York, R. (1966). *Equality of Educational Opportunity*. Washington, D.C.: National Centre for Educational Statistics.

Collins, L. and Horn, J. (Eds.)(1991). *Best Methods for the Analysis of Change*. Washington, D.C.: APA.

Coombs, J.R. (1994). Equal Access to Education: the Ideal and the Issues. *Journal of Curriculum Studies* 26(3), 281-295.

Cooper, D., and Subotzky, G. (2001). *The Skewed Revolution: Trends in South African Higher Education*. Bellville, South Africa: Educational Policy Unit, University of the Western Cape.

Craig, A.P. (1991). Adult Cognition and Tertiary Studies. *South African Journal for Higher Education* 5(2),137-144.

Criper, C. and Davies, A. (1988). *Research Report 1(i) ELTS Validation Project Report*. Cambridge: British Council/University of Cambridge Local Examinations Syndicate.

Cronbach, L.J. (1988). Five Perspectives on Validity Argument. In Wainer, H. (Ed.)(1988). *Test Validity*. Hillsdale, NJ.: Erlbaum, 3-17.

Cronbach, L.J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 297-334.

Crouch, L. (1999). Education Data and the 1996 Census: Some Crucial Apparent Problems and Possible Strategies for Resolution. *EduSource 24,* 10-16.

Crouch, L. and Mabogoane, T. (1997). Aspects of Internal Efficiency Indicators in South African Schools: Analysis of Historical and Current Data. *EduSource* 19, 4-28.

Crouse, J. and Trusheim, D. (1988). *The Case Against the SAT*. Chicago: University of Chicago Press.

Cumming, A., Kantor, R., Powers, D., Santos, T. and Taylor, C. (2000). *TOEFL 2000 Writing Framework: A Working Paper*. TOEFL Monograph Series. Princeton NJ: Educational Testing Service.

Cummins, J. (2000). *Language, Power and Pedagogy: Bilingual Children in the Crossfire*. Clevedon: Multilingual Matters Ltd.

Cummins, J. (1984). Implications of Bilingual Proficiency for the Education of Minority Language Students. *Language Issues and Education Policies*. ELT Documents 119. Oxford: Pergamon Press and The British Council.

Cummins, J. (1980). The Cross-Lingual Dimensions of Language Proficiency: Implications for Bilingual Education and the Optimal Age Issue. *TESOL Quarterly* 14, 175-87.

Cummins, J. and Swain, M. (1986). *Bilingualism in Education*. New York: Longman.

Davies, A. (1988). Operationalising Uncertainty in Language Testing: An Argument in Favour of Content Validity. *Language Testing* 5(1), 32-48.

Davies, A. (1985). *Communicative Language Testing: A Sceptical View*. Paper presented at British Council Course 559: Communicative Language Testing, University of Lancaster, Lancs.: United Kingdom.

Dawes, P., Yeld, N. and Smith, M.J. (1999). Access, Selection and Admission to Higher Education: Maximising the Use of the School-Leaving Examination. *South African Journal of Higher Education* 13(3), 97-104.

De Groot, A. (1965). *Thought and Choice in Chess*. The Hague: Mouton.

Department of Education (2001). *National Plan for Higher Education*. Pretoria: South Africa.

Department of Education (2000). *A South African Curriculum for the Twenty-First Century*. Report of the Review Committee on Curriculum 2005. Pretoria: South Africa.

Department of Education (1998). *Investigation into the Senior Certificate Examination*. Report of the Ministerial Committee on the Senior Certificate Examination. Pretoria: South Africa.

Department of Education (1997a). *Curriculum 2005. Lifelong Learning for the 21st Century*. Pretoria: South Africa.

Department of Education (1997b). *Education White Paper 3. A Programme for the Transformation of Higher Education*. General Notice 1196 in the *Government Gazette* 18207 of 1997-08-15. Pretoria: South Africa.

Department of Education (1997c). Language in Education Policy. *Government Gazette* 17997(383). Pretoria: Government Printer.

Department of Education (1996). *Green Paper on Higher Education Transformation*. Pretoria: South Africa.

Department of Education and Culture (1995). *Guidelines for the Implementation of the New ESL Interim Core Syllabuses (English Second Language Higher Grade, Senior Secondary Phase).* Pretoria: Department of Education.

Department of Education and Culture (1985). *Syllabus for English Second Language Higher Grade, Standards 9 and 10*. Pretoria: Department of Education and Culture.

Donlon, T.F. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Draper, J.A. (1994). Looking at Philosophies for Literacy Education. In Draper, J.A. and Taylor, M.C. (Eds.)(1994). Voices from the Literacy Field. Malabar, Fla.: Krieger Publishing Company, 3-20.

Draper, J.A. and Taylor, M.C. (Eds.)(1994). *Voices from the Literacy Field*. Malabar, Fla.: Krieger Publishing Company.

Drenth, P.J.D., Van der Flier, H. and Omari, I.M. (1983). Educational Selection in Tanzania. *Evaluation in Education* 7, 93-217.

Eckstein, M.A. and Noah, J.N. (Eds.)(1993). *Secondary School Examinations: International Perspectives on Policies and Practice.* Ann Arbor, Michigan: Yale University Press.

Educational Testing Service (2000). *ETS Standards for Quality and Fairness*. Princeton, NJ.: Educational Testing Service.

Educational Testing Service (1999). *ETS Strivers Study: Background.* URL: http://144.81.21.133/access/Stories-FYI/construct.htm (accessed 8 October 1999).

Educational Testing Service Trustees' Colloquy (1995). *Performance Assessment: Different Needs, Difficult Answers*. Princeton NJ: Educational Testing Service.

Ekstein, M.A. and Noah, J.N. (Eds.)(1992). *Examinations: Comparative and International Studies.* Exeter: Pergamon Press.

Embretson, S.E. (1992). Measuring and Validating Cognitive Modifiability as an Ability: A Study in the Spatial Domain., *Journal of Educational Measurement* 29(1), 25-50.

Enright, M.K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. and Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper.* TOEFL Monograph Series. Princeton NJ.: Educational Testing Service.

Flesch, R.F. (1948). A New Readability Yardstick. *Journal of Applied Psychology* 32, 221-33.

Foster, P.J. (1992). Commentary. In Ekstein M.A. and Noah, J.N. (Eds.)(1992). *Examinations: Comparative and International Studies.* Exeter: Pergamon Press, 121-126.

Frederiksen, N. (1994). The Influence of Minimum Competency Tests on Teaching and Learning. Princeton, NJ.: Educational Testing Service.

Freebody, P. (1997). Assessment as Communal Versus Punitive Practice: Six New Literacy Crises. Virtual Seminar 1, International Association of Applied Linguistics. URL: http://www.education.uts.edu.au/AILA/VirtSem.Free (accessed 22 June 1999).

Freebody, P. and Welch, A.R. (Eds.)(1993). Knowledge, Culture and Power: International Perspectives on Literary Policies and Practices. London: Falmer Press.

Garcia, E.E., Jorgensen, R.E., and Ormsby, C. (1999). How Can Public Universities Still Admit a Diverse Freshman Class? The Case of Latinos, the SAT, and the University of California. *The Journal of College Admission* Summer/Fall, 5-11.

Gee, J.P. (1990). *Social Linguistics and Literacies: Ideology In Discourses.* Basingstoke: The Falmer Press.

Gee, J.P. (2000). The New Literacy Studies. From 'Socially Situated' to the Work of the Social. In Barton, D., Hamilton, M. and Ivanic, R. (Eds.)(2000). *Situated Literacies: Reading and Writing in Context.* London: Routledge, 180-196.

Geisinger, K.F. (1992). The Metamorphosis of Test Validation. *Educational Psychologist* 27(2), 197-222.

Gifford, B.R. and O'Connor, M.C. (Eds.)(1992). *Changing Assessments. Alternative Views of Aptitude, Achievement and Instruction.* Boston: Kluwer Academic Publishers.

Gipps C. and Murphy, P. (1994). *A Fair Test? Assessment, Achievement and Equity.* Buckingham: Open University Press.

Glaser, R. (1991). Expertise and Assessment. In Wittrock, M.C. and Baker, E.L. (Eds.)(1991). *Testing and Cognition.* New Jersey: Prentice Hall, 17-30.

Glaser, R. (Ed.)(1982). Advances in Instructional Psychology, Volume 2. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldman, S.R. (1997). Learning from Text: Reflections on the Past and Suggestions for the Future. Discourse Processes 23, 357-398.

Goldman, S.R. and Saul, E.U. (1990). Flexibility in Text Processing: A Strategy Competition Model. Learning and Individual Differences 2, 181-219.

Goody, J. (Ed.) (1968). Literacy in Traditional Societies. Cambridge: Cambridge University Press.

Green, B.F. (1995). Setting Performance Standards: Content, Goals and Individual Differences. William H Angoff Memorial Lecture Series. Princeton NJ.: Educational Testing Service.

Hamers, J.H.M., Sijtsma, K., and Ruijssenaars, A.J.J.M. (Eds.)(1993). *Learning Potential Assessment. Theoretical, Methodological and Practical Issues.* Amsterdam/Lisse: Swets and Zeitlinger B.V.

Hamilton, J., Lopes, M., McNamara, T.F., and Sheridan, E. (1993). Rating Scales and Native Speaker Performance on a Communicatively Oriented EAP Test. *Language Testing* 10(3), 337-353.

Hamp-Lyons, L. and Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, Community and Assessment.* TOEFL Monograph Series Report No. 5. Princeton, NJ.: Educational Testing Service.

Hansen, J. (1997). *A Comparative Documentary Analysis of ESL Matriculation Examination Papers (1994-1996) within the Context of the Shift from a Segregated to a Unitary Education System.* Unpublished Minor MPhil Dissertation. Cape Town: University of Cape Town.

Harley, B., Allen, P., Cummins, J. and Swain, M. (1990). *The Development of Second Language Proficiency.* Cambridge: Cambridge University Press.

Hamilton, J. (1991). *Native and Non-Native Speaker Performance on the IELTS Reading Test.* Unpublished MA Thesis, University of Melbourne.

Hatch, E. and Farhady, H. (1982). *Research Design and Statistics for Applied Linguistics.* Rowley, Mass.: Newbury House.

Heath, S.B. (1980). The Functions and Uses of Literacy. *Journal of Communications* 30, 123-133.

Herman, H.D. (1995). School-Leaving Examinations, Selection and Equity in Higher Education in South Africa. *Comparative Education* 31(2), 261-274.

Heubert, J.P. and Hauser, R.M. (1999). *High Stakes. Testing for Tracking, Promotion and Graduation.* Washington, DC: National Academy Press.

Heyneman, S.P. and Ransom, A.W. (1990). Using Examinations and Testing to Improve Educational Quality. In Ekstein, M.A. and Noah, J.N. (Eds.)(1992). *Examinations: Comparative and International Studies.* Exeter: Pergamon Press, 105-120.

Hidi, S. and Anderson, V. (1986). Producing Written Summaries: Task Demands, Cognitive Operations, and Implications for Instruction. *Review of Educational Research* 56(4), 473-493.

Hill, C. and Parry, K. (1994). Models of Literacy: the Nature of Reading Tests. In Hill, C. and Parry, K. (Eds.)(1994). *From Testing to Assessment. English as an International Language.* New York: Longman, 7-34.

Hill, C. and Parry, K. (Eds.)(1994). *From Testing to Assessment. English as an International Language.* New York: Longman.

Hill, K., Storch, N. and Lynch, B. (1999). A Comparison of IELTS and TOEFL as Predictors of Academic Success. In Tulloh, R. (Ed.)(1999). *IELTS Research Reports*, Vol.2, IELTS Australia Pty. Ltd., 52-63.

Kellaghan, T. and Greaney, V. (1992). *Using Examinations to Improve Education. A Study in Fourteen African Countries.* World Bank Technical Paper No. 165, Africa Technical Department Series. Washington: The World Bank.

Kellaghan,T., Madaus, G.F, and Raczek, A. (1996*). The Use of External Examinations to Improve Student Motivation*. A Public Service Monograph of the American Educational Research Association.

Kintsch, W. (1988). The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. *Psychological Review* 95, 163-182.

Kirsch, I.S., Jungeblat, A., and Campbell, A. (1992). *Beyond the School Doors. The Literacy Needs of Job Seekers Served by the U.S. Department of Labor.* Princeton, NJ.: Educational Testing Service.

Klein, S.P., Josavnoic, J., Stecher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G. and Comfort, K. (1997). Gender and Racial/Ethnic Differences on Performance Assessments in Science. *Educational Evaluation and Policy Analysis* 19(2), 83-97.

Klitgaard, R. (1985). *Choosing Elites. Selecting the "Best and the Brightest" at Top Universities and Elsewhere.* New York: Basic Books Inc.

Koretz, D., Linn, R.L., Dunbar, S.B., and Shepard, L.A. (1991). *The Effects of High-stakes Testing on Achievement: Preliminary Findings about Generalization across Tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Labov, W. (1973). The Logic of Nonstandard English. In Keddie, N. (Ed.)(1973). *Tinker, Tailor, the Myth of Cultural Deprivation.* Harmondsworth: Penguin Books, 21-66.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests: A Teacher's Book.* New York, NY.: McGraw-Hill.

Lado, R. (1946). *Test of Aural Comprehension in English.* Ann Arbor, Mich.: English Language Institute, University of Michigan.

Langer, J.A., Applebee, A.N., Mullis, I.V. and Foerthsch, A. (1990). *Learning to Read in our Nation's Schools: Instruction and Achievement in 1988 at Grades 4, 8 and 12.* Princeton, NJ: Educational Testing Service.

Lave, J. (1996). Teaching, as Learning, in Practice. *Mind, Culture and Activity* 3: 149-164.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: John Wiley and Sons.

Lazaraton, A., Riggenbach, H. and Ediger, A. (1987). Forming a Discipline: Applied Linguists' Literacy in Research Methodology and Statistics. *TESOL Quarterly* 21(2), 236-page.

Lee, Y.P., Fok, A.C.Y.Y., Lord, R. and Low, G. (Eds.)(1985). *New Directions in Language Testing.* Oxford: Pergamon Press.

Lemann, N. (1999a). Behind the SAT. *Newsweek* September 6, 1999, 52-57.

Lemann, N. (1999b). *The Big Test: The Secret History of the American Meritocracy.* New York, NY.: Farrar, Straus and Giroux.

Lentolf, J.P. and Labares, A. (Eds.)(1987). *Research in Second Language Learning: Focus on the Classroom.* Norwood N.J.: Ablex.

Leont'ev, A.N. (1978). *Activity, Consciousness, and Personality.* Englewood Cliffs, NJ.: Prentice-Hall.

Lidz, C.S. (Ed.)(1987a). *Dynamic Assessment. An Interactional Approach to Evaluating Learning Potential.* New York: The Guilford Press.

Lidz, C.S. (1987b). Cognitive Deficiencies Revisited. In Lidz, C.S. (Ed.)(1987). *Dynamic Assessment. An Interactional Approach to Evaluating Learning Potential.* New York: The Guilford Press, 444-475.

Lin, A.M.Y. (1999). Doing-English-Lessons in the Reproduction or Transformation of Social Worlds? *TESOL Quarterly* 33(3), 393-412.

Lindquist, E.F. (Ed.)(1951). *Educational Measurement.* Washington, DC.: American Council on Education.

Linn, R.L. (Ed.)(1989). *Educational Measurement* (3rd Edition). New York: ACE and Macmillan.

Linn, R.L., Baker, E.L, and Dunbar, S.B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher* 20(8), 15-21.

Lopes, M. (1992). *Native Speaker Performance on the IELTS Reading Test.* Unpublished MA Thesis, University of Melbourne.

Lotter C. (1994). Determining Matriculation Requirements by Virtue of the Senior Certificate Examination Results: Matriculation Board Perspective. In Barsby, T., Haeck, W. and Yeld, N. (Eds.) (1994). *Accountability in Testing: the Contribution of Tests to Increased Access to Post Secondary Education.* South African Association for Academic Development - Workshop Publication Series No.2. Cape Town: University of Cape Town, 21 - 23 and 39 - 49.

Loubser, N.D. (1997). Joint Selection Programme for Science and Applied Science (JSPSAS): Statistical Analysis of 1997 Intake Data. Unpublished Manuscript, University of Natal.

MacDonald, C.A. (1990). *Crossing the Threshold into Standard Three.* Pretoria: Human Sciences Research Council.

Madaus,G.F. (1988). The Influence of Testing on the Curriculum. In Tanner L.N. (Ed.)(1988). *Critical Issues in Curriculum.* Chicago: University of Chicago Press, 83-121.

Masters, G.N. and Forster, M. (1997). *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey.* Canberra: Department of Employment, Education, Training and Youth Affairs.

Mayer, R.E. (1989). Teaching for Thinking: Research on the Teachability of Thinking Skills. In Cohen, I.S. (Ed.)(1989). *The G. Stanley Hall Lecture Series* 9. Washington, DC.: American Psychological Association.

McNamara, T.F. (1996). *Measuring Second Language Performance*. New York: Longman.

Messick, S. (1989). Validity. In Linn, R.L.(Ed.)(1989). *Educational Measurement* (3rd Edition). New York: ACE and Macmillan, 13-103.

Messick, S. (Ed.)(1999). *Assessment in Higher Education. Issues of Access, Quality, Student Development, and Public Policy.* Matwah, NJ.: Lawrence Erlbaum.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher* 23(2), 13-23.

Miller, R. (1989). Conceptual Issues in Theorising about Cognition. *South African Journal for Higher Education* 3(1), 154-159.

Miller, R. (1992). Double Double Toil and Trouble: The Problem of Student Selection. *South African Journal for Higher Education* 6(1), 98-104.

Mitchell, G., Fridjhon, P., and Haupt, J. (1997). On the Relationship between the Matriculation Examination and University Performance in South Africa - 1980 to 1991. *South African Journal of Science* 93, 382-387.

Moll, I. and Slonimsky, L. (1989). Towards an Understanding of Cognition and Learning in the Academic Support Context. *South African Journal of Higher Education* 3(1), 160-166.

Monare, M. (2001). 'Impossible' Matric-Pass Rise to be Probed. *The Star* January 17, 2001.

Morris, A. (1985). *Schooling, Culture and Class: A Study of White and Coloured Schooling and its Relationship to Performance in Sociology at the University of Cape Town.* Unpublished MSocSc Thesis. Cape Town: University of Cape Town.

Morrow, K. (1981). Communicative Language Testing: Revolution or Evolution? In Alderson, J.C. and Hughes, A. (Eds.)(1991). *ELT Documents 111 – Issues in Language Testing.* London: The British Council, 9-25.

Morrow, W. (1994). Entitlement and Achievement in Education. *Studies in Philosophy and Education* 13, 33-47.

Moses, Y.T. (1994). Quality, Excellence, and Diversity. In Smith, D.G., Wolf, L.E. and Levitan (Eds.)(1994). *Studying Diversity in Higher Education.* San Francisco: Jossey-Bass.

Muller, J. and Roberts, J. (2000). *The Sound and Fury of International School Reform: A Critical Review.* Report Prepared for the Joint Education Trust, February 2000. Pretoria: Joint Education Trust.

Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Mussen, P. (Ed.)(1983). *Handbook of Child Psychology (4th Edition), Vol. 1: History, Theory and Methods.* New York: John Wiley.

Naidu, E. (2000). Student Enrolment Expected to Drop. *The Star* 6 February, 2000.

National Academy of Education (1997). *Assessment in Transition: Monitoring the Nation's Educational Progress*. Washington DC: National Academy Press.

National Commission on Excellence in Education (1983). *A Nation at Risk: The Imperative for Educational Reform*. Washington, DC: National Commission on Excellence in Education.

Ndebele, N.S. Maintaining Domination Through Language. *Academic Development* 1(1), 3-5.

Nitko, A.J. (2001). *Educational Assessment of Students (3rd Edition)*. Upper Saddle River, NJ.: Merrill Prentice Hall.

Noah H.J. and Eckstein, M.A. (1992). Comparing National Systems of Secondary School-Leaving Examinations. In Eckstein and Noah (Eds.) (1992). *Examinations: Comparative and International Studies*. Exeter: Pergamon Press, 3-24.

Norton Peirce, B. (1992). Demystifying the TOEFL Reading Test. *TESOL Quarterly* 26(4), 665-689.

Noss, R., Goldstein, H., and Hoyles, C. (1989). Graded Assessment and Learning Hierarchies in Mathematics. *British Educational Research Journal* 15(2), 109-120.

Office of Educational Assessment, University of Washington, Wa. (1998). Reliability. URL: http://www.washington.edu/ oea/testhome.htm (Accessed 9 January 2001).

Ogbu, J.U. (1982). Equalisation of Educational Opportunity and Racial/Ethnic Inequality. In Altbach, P.G., Arnove, R.F., and Kelly, G.P. (Eds.)(1982). *Comparative Education*. New York: MacMillan, 269-289.

Oller, J.W. (Ed.)(1983). *Issues in Language Testing Research*. Rowley, MA: Newbury House.

Oller, J.W. (1981). Language Testing Research 1979-1980. In Kaplan, R., Jones, R.L. and Tucker, G.R. (Eds.)(1981). *Annual Review of Applied Linguistics – Volume 1*. Rowley, MA.: Newbury House, 124-150.

Oller, J.W. (1979). *Language Tests at School: A Pragmatic Approach*. Longman.

Oller, J.W. and Conrad, C. (1971). The Cloze Technique and ESL Proficiency. *Language Learning* 21(2), 183-195.

Oller, J.W. and Perkins, K. (1980). *Research in Language Testing*. Rowley, MA.: Newbury House.

Olson, D.R. (1977). From Utterance to Text: The Bias of Language in Speech and Writing. *Harvard Educational Review* 47(3), 257-281.

Pahad, M.C. (1996). Issues of Assessment in Education in South Africa during the Period of Transition 1989-1995. Unpublished MPhil Dissertation. Johannesburg: University of the Witwatersrand.

Palinscar, A.S. (1986). The Role of Dialogue in Providing Scaffolded Instruction. *Educational Psychologist* 21(1&2), 73-98.

Palinscar, A.S. and Brown, A.L. (1984). Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities. *Cognition and Instruction* 1(2), 117-175.

Pellegrino, J.W., Jones, L. R., and Mitchell, K.J. (Eds.)(1999). *Grading the Nation's Report Card. Evaluating NAEP and Transforming the Assessment of Educational Progress.* Washington: National Academy Press.

Pennycook, A. (1994). *The Cultural Politics of English as an International Language*. London: Longman.

Perfetti, C.A. (1989). There are Generalized Abilities and One of Them is Reading. In Resnick, L.B. (Ed.)(1989). *Knowing, Learning and Instruction: Essays in Honor of Robert Glaser*. Hillsdale, NJ.: Erlbaum, 307-333.

Plüddemann, P., Mati, X. and Mahlalela-Thusi, B. (1999). *Problems and Possibilities in Multilingual Classrooms in the Western Cape.* Unpublished Report on a Study Conducted for the Project for the Study of Alternative Education in South Africa (PRAESA). Cape Town: University of Cape Town.

Polakow, D. (1999). *Survival-Analysis of University Tenure, with Foci on Contrasting Probability of Exclusion in AARP and non-AARP ex-DET Student Groupings.* Report Prepared for Discussion by the Academic Development Programme. Cape Town: University of Cape Town.

Polakow, D. (1998a). *Analysis of the Predictive Capacity of AARP Testing: The 1995 Cohort. Report Prepared for Discussion by the Academic Development Programme*. Cape Town: University of Cape Town.

Polakow, D. (1998b). *Analysis of the Predictive Capacity of AARP Testing: The 1996 Cohort – Second Year. Report Prepared for Discussion by the Academic Development Programme*. Cape Town: University of Cape Town.

Polakow, D. (1997). *Analysis of the Predictive Capacity of AARP Testing: The 1996 Cohort. Report Prepared for Discussion by the Academic Development Programme.* Cape Town: University of Cape Town.

Potter, C.S. and Jamotte, A.N. (1985). African Matric Results. Dubious Indicators of Academic Merit. *Indicator South Africa* 3(1), 10-13.

Price, S. (1999). Critical Discourse Analysis: Discourse Acquisition and Discourse Practices. *TESOL Quarterly* 33(3), 581-595.

Pride, J.B. and Holmes, J. (Eds.)(1972). *Sociolinguistics: Selected Readings.* Harmondsworth: Penguin.

Prinsloo, M. and Breier, M. (Eds.)(1996). *The Social Uses of Literacy: Theory and Practice in Contemporary South Africa*. Johannesburg: Sached and John Benjamin Publishing Company.

Queensland Board of Senior Secondary School Studies (1999). *Cross-Curriculum Testing*, URL: http://bssssq.Edu.au/Assessment/Cross-curriculumTesting.html (accessed 13 November 1999).

Raimes, A. (1991). Out of the Woods: Emerging Traditions in the Teaching of Writing. *TESOL Quarterly* 25(3), 407-430.

Report Cites Racial Gap in Student Performance. *The Chronicle of Higher Education* XLV1(10), A42. October 29, 1999.

Resnick, L.B. (Ed.)(1989). *Knowing, Learning and Instruction: Essays in Honor of Robert Glaser*. Hillsdale, NJ:Erlbaum, page .

Resnick, L.B. and Resnick, D.P. (1992). Assessing the Thinking Curriculum: New Tools for Educational Reform. In Gifford, B.R. and O'Connor, M.C. (Eds.)(1992). *Changing Assessments. Alternative Views of Aptitude, Achievement and Instruction*. Boston: Kluwer Academic Publishers, 37-75.

Richards, J.C., and Schmidt, R.W. (Eds.)(1983). *Language and Communication*. London: Longman.

Richardson, K. and Spears, D. (Eds.)(1972). *Race, Culture and Intelligence*. Harmondsworth: Penguin.

Robbins, D. (1999). Tertiary Education: The Size and Shape of Things to Come. *Daily Mail and Guardian* November 24.

Rodseth, V. (1995). *Bilingualism and Multilingualism in Education*. Johannesburg: Centre for Continuing Development, University of the Witwatersrand.

Rogoff, B. and Lave, J. (Eds.)(1984). *Everyday Cognition*. Cambridge, MA.: Harvard University Press

Rogoff, B. (1984). Introduction: Thinking and Learning in Social Context. In Rogoff, B. and Lave, J. (Eds.)(1984). *Everyday Cognition*. Cambridge, MA.: Harvard University Press, 1-8.

Ryan, J. (1972). The Illusion of Objectivity. In Richardson, K. and Spears, D. (Eds.)(1972). *Race, Culture and Intelligence*. Harmondsworth: Penguin.

Ryans, D.G. and Frederiksen, N. (1951). Performance Tests of Educational Achievement. In Lindquist, E.F. (Ed.)(1951). *Educational Measurement*. Washington, DC.: American Council on Education, 455-491.

Segall, R.L. (1994). University Access and Admissions Policy. In Barsby, T., Haeck, W. and Yeld, N. (Eds.) (1994). *Accountability in Testing: the Contribution of Tests to Increased Access to Post Secondary Education*. South African Association for Academic Development - Workshop Publication Series No.2. Cape Town: University of Cape Town, 4-12.

Sewell, T.E. (1979). Intelligence and Learning Tasks as Predictors of Scholastic Achievement in Black and White First-Grade Children. *Journal of School Psychology* 17, 325-332.

Taylor, N. and Vinjevold, P. (Eds.)(1999). Getting Learning Right. Report of the President's Education Initiative Research Project. Johannesburg: Joint Education Trust.

Taylor, N. and Vinjevold, P. (1999). Teaching and Learning in South African Schools. In Taylor, N. and Vinjevold, P. (Eds.)(1999). Getting Learning Right. Report of the President's Education Initiative Research Project. Johannesburg: Joint Education Trust, 131-162.

The College Board (1999a). *Facts About the SAT 1: Reasoning Test.* Information Sheet issued by the College Board on the Educational Testing Service '*Infoline*", October 1999.

The College Board (1999b). *Concordance Between SAT1 and ACT Scores for Individual Students.* Research Notes RN-07. New York: The College Board.

The College Board (1997). *Common Sense About SAT Score Differences and Test Validity*. Research Notes, RN-01, New York, NY.: The College Board.

The College Board (1990). *Beyond Prediction*. New York, NY.: The College Board.

Thernstrom, S. and Thernstrom, A. (1997). *America in Black and White: One Nation Indivisible*. New York; Simon and Shuster.

Thesen, L. (1997). Voices, Discourse and Transition: In Search of New Categories. *TESOL Quarterly* 31(3): 487-511.

Tierney, W.G. (1997). The Parameters of Affirmative Action: Equity and Excellence in the Academy. *Review of Educational Research* 45, 89-125.

Ting, S.R. and Robinson, T.L. (1998). First-Year Academic Success: A Prediction Combining Cognitive and Psychosocial Variables for Caucasian and African American Students. *Journal of College Student Development* 39(6), 599 – 610.

Tusting, K., Ivanic, R., and Wilson, A. (2000). New Literacy Studies at the Interchange. In Barton, D., Hamilton, M. and Ivanic, R. (Eds.)(2000). *Situated Literacies: Reading and Writing in Context*. London: Routledge, 210-218.

University of Cape Town (2001). An Analysis of the 2001 Admissions Cycle. Cape Town: Academic Planning Office, University of Cape Town.

University of Cape Town (undated). Proposal for a Research Project to Devise Alternative Selection Procedures for Entry to UCT. Cape Town: Academic Planning Office, University of Cape Town.

Van den Berg, J. (1989). Intelligence Tests. In Owen, K. and Taljaard, J.J. (Eds.)(1989). *Handbook for the Use of Psychological and Scholastic Tests of IPER and the NIPR*. Pretoria: Human Sciences Research Council, 83-135.

Vinjevold, P. (1999). Language Issues in South African Classrooms. In Taylor, N. and Vinjevold, P. (Eds.)(1999). Getting Learning Right. Report of the President's Education Initiative Research Project. Johannesburg: Joint Education Trust, 205-226.

Vygotsky, L.S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wagner, L (1989). Access and Standards: An Unresolved (and Unresolvable?) Debate. In Ball, C. and Eggins, H. (Eds.)(1989). *Higher Education into the 1990s: New Dimensions*. Milton Keynes: Society for Research into Higher Education and Open University Press, 29-37.

Wainer, H. (Ed.)(1988). *Test Validity*. Hillsdale, NJ.: Erlbaum

Waters, A. (1996). *A Review of Research into Needs in English for Academic Purposes of Relevance to the North American Higher Education Context*. Princeton, NJ.: Educational Testing Service.

Weir, C.J. (1990). Communicative Language Testing. Englewood Cliffs NJ.: Prentice-Hall Regent.

Weir, C.J. (1988). Construct Validity. In Hughes, A., Porter, D. and Weir, C. (Eds.)(1988). ELTS Validation Project: Proceedings of a Conference ON the ELTS Validation Project Report. English Language Testing Service Research Report 1(ii). London: British Council/University of Cambridge Local Examinations Syndicate, 15-25.

Weir, C.J. (1983). Identifying the Language Needs of Overseas Students in Tertiary Education in the United Kingdom. Unpublished PhD Thesis, University of London Institute of Education.

Weissert, W. (1999). Report Cites Racial Gap in Student Performance. Chronicle of Higher Education, 29 Oct. 1999: A42).

Welch, A.R. and Freebody, P. (1993). Crisis and Context in Literacy Education. In Freebody, P. and Welch, A.R. (Eds.)(1993). Knowledge, Culture and Power: International Perspectives on Literary Policies and Practices. London: Falmer Press.

Wells, G. (Ed.)(1981). *Learning Through Interaction. The Study of Language Development*. Cambridge: Cambridge University Press.

Whitehead, A.N. (1929). *The Aims of Education*. New York: MacMillan.

Widdowson, H.G. (1998a). Context, Community, and Authentic Language. *TESOL Quarterly* 32(4), 705-715.

Widdowson, H.G. (1998b). The Theory and Practice of Critical Discourse Analysis. *Applied Linguistics* 19(1), 136-151.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity. *Phi Delta Kappan* 75(3), 200-214.

Willett, J.B. and Singer, J.D. (1991). How Long Did It Take? Using Survival Analysis in Educational and Psychological Research. In Collins, L. and Horn, J. (Eds.)(1991). *Best Methods for the Analysis of Change*. Washington, D.C.: APA, 310-348.

Wittrock, M.C. (1991). Testing and Recent Research in Cognition. In Wittrock, M.C. and Baker, E.L. (Eds.)(1991). *Testing and Cognition*. New Jersey: Prentice Hall, 4-16.

Wittrock, M.C. and Baker, E.L. (Eds.)(1991). *Testing and Cognition*. New Jersey: Prentice Hall.

Woodrow, M. (1986). Scrutiny in Partnership: Access Issues from Eighteen Courses in South London. *Journal of Access Studies* 1(1), 33-43.

Yardley, J. (2000). A Test is Born. *New York Times* Section 4A, 9 April 2000.

Yeld, N. (2000). *The Development of Version 3 of the TELP Placement Test in Academic Literacy*. Unpublished Report for the Desmond Tutu Educational Trust, October 2000.

Yeld, N. (1999). *Tertiary Education Linkages Project II: Report on the Development of the Standardised Tests*. Unpublished Report for the Desmond Tutu Educational Trust, April 1999.

Yeld, N. (1995). *The School-Leaving Examination as a Regulatory Mechanism for Access to Post-Secondary Education: Possible Developments and Consequences*. Paper Commissioned by the Access and Admissions Task Group of the National Commission on Higher Education, 1995.

Yeld, N. (1987). *Communicative Language Testing and Validity*. Unpublished Course Paper, Presented in Partial Fulfillment of the Requirements for the Degree of Master of Education. Cape Town: University of Cape Town, 1-26.

Yeld, N., Badsha, N. and Shall, A. (1989). *English Language Proficiency Test: Progress Report*. Unpublished Manuscript, Alternative Admissions Research Project, University of Cape Town.

Yeld, N., Clark, S., Davies, N., Day, J., de Boer, A., Hoepner, L., Inglis, M., Kaunda, L., Shabalala, L., Thesen, L., and Visser, A. (2000). *Placement Test in English for Educational Purposes: Radio*. Cape Town: University of Cape Town.

Yeld, N., Clark, S., Hansen, J., Hartman, N., Hoepner, L., Inglis, M., Pinto, D., Shabalala, L., and Thesen, L. (1999). *Placement Test in English for Educational Purposes: Water*. Cape Town: University of Cape Town.

Yeld, N., Hansen, J., Hartman, Inglis, M., Pinto, D., and Thesen, L. (1998). *Placement Test in English for Educational Purposes: Fire*. Cape Town: University of Cape Town.

Yeld, N., Hansen, J., Hartman, N., Inglis, M., Pinto, D., and Thesen, L. (1997). *Placement Test in English for Educational Purposes: Antarctic*. Cape Town: University of Cape Town.

Yeld, N. and Haeck, W. (1997). Educational Histories and Academic Potential: Can Tests Deliver? *Assessment and Evaluation in Higher Education* 22, 5 -16.3.

Yeld, N. and Hartman, N. (1992). Tasks, Performances and Placement: Implications for Selection and Educational Intervention. In Badsha, N., Griesel, H., Smith, M. and Yeld, N. (Eds.)(1992). Proceedings of Cintsa Admissions Symposium, 42-58.

Yeld, N. and van Bommel, L. (1997). Are We Playing Games with the Matric? *Cape Argus*, 17 April 1997.

Yeld, N. and Visser, A. (1999). Alternative Admissions Research Project: 1999 Annual Report. AARP Internal Report, February 2000.

Zaaiman, H. (1998). *Selecting Students for Mathematics and Science. The Challenge Facing Higher Education in South Africa*. Pretoria: Human Sciences Research Council.

Zaaiman, H. (1996). *The 1994 UNIFY Student Group Tracer Study Report*. UNIFY Internal Report, August 1996.